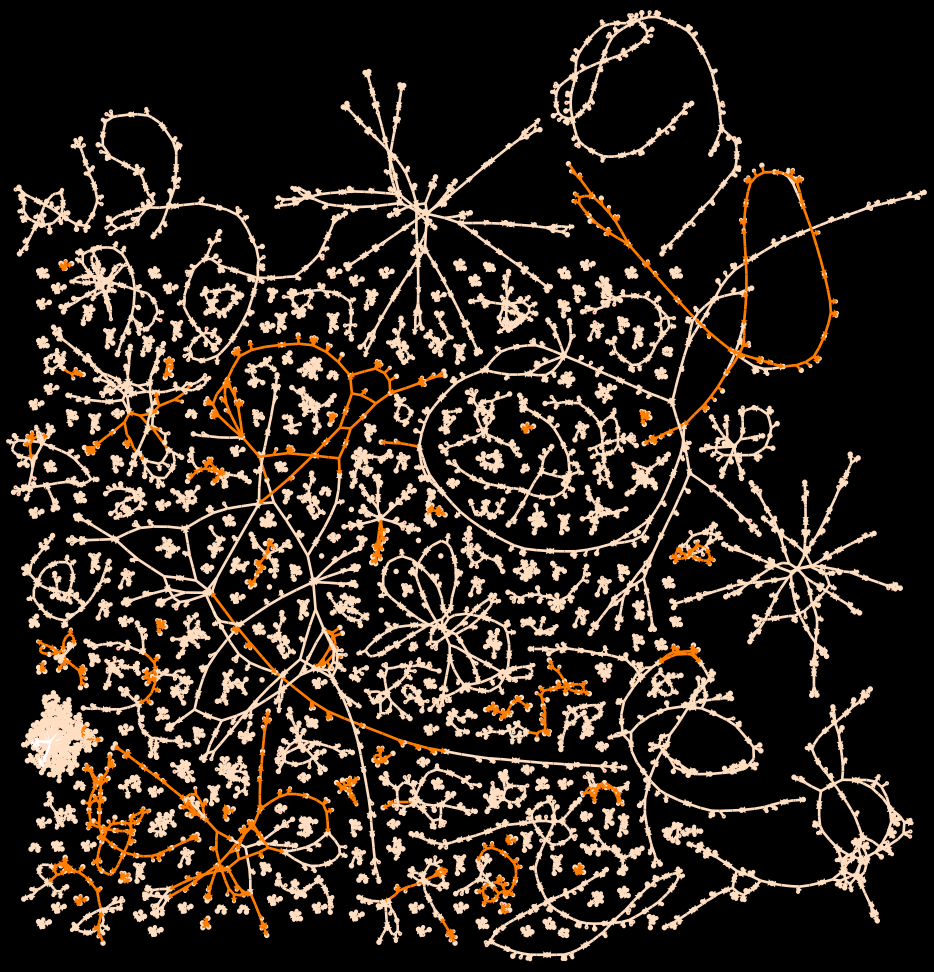
Metabolic adaptation of a human pathogen during chronic infections - a systems biology approach

Thøgersen, Juliane Charlotte

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

METABOLIC ADAPTATION OF A HUMAN PATHOGEN DURING CHRONIC INFECTIONS - A SYSTEMS BIOLOGY APPROACH



PhD Thesis by

Juliane Charlotte Thøgersen

Infection Microbiology Group
Department of Systems Biology
Technical University of Denmark

DTU

March 2015

METABOLIC ADAPTATION OF A HUMAN PATHOGEN DURING CHRONIC INFECTIONS
- A SYSTEMS BIOLOGY APPROACH

Infection Microbiology Group
Department of Systems Biology
Technical University of Denmark

Cover illustration: Map of *Pseudomonas aeruginosa* metabolism highlighting altered reaction activities between *P. aeruginosa* isolates from initial and chronic infections in cystic fibrosis patients. Developed by Juliane C. Thøgersen and Jennifer A. Bartell using the software MetDraw (Jensen & Papin, 2014).

"All models are wrong.

Some are useful."

George Box (1979)

Preface

This thesis is written as a partial fulfillment of the requirements to obtain a PhD degree at the Technical University of Denmark (DTU). The work presented in this thesis was carried out from August 2011 to March 2015 at the Infection Microbiology Group (IMG), Department of Systems Biology, DTU, under the supervision of Professor Søren Molin and Associate Professor Lars Jelsbak. From September 2014 to December 2014 the work was carried out at the CfB group at the Novo Nordisk Foundation Center for Biosustainability, DTU. Part of the work was carried out in close collaboration with Associate Professor Jason Papin and Jennie Bartell from University of Virginia, USA. This PhD project was financed by the Danish Council for Independent Research.

Juliane Charlotte Thøgersen

Kgs. Lyngby, March 2015

Acknowledgements

First of all, I want to thank my supervisors Søren Molin and Lars Jelsbak for their guidance and support during countless scientific discussions. Thank you for always showing a great trust in me, allowing me to work independently, but still with your full engagement when needed. This trust has been the main driver and motivator during my PhD work. Søren, I am very grateful for your role as a mentor during my Master's study, which was crucial for my decision to obtain a PhD. Your strong enthusiasm for research has always inspired me and will continue to do so throughout my career.

I also wish to express my greatest gratitude to my collaborators Jette Thykær and Kristian F. Nielsen, Morten Mørup, and Jason A. Papin and Jennifer A. Bartell. Especially thanks to Jennie for being a great teammate and for our frequent scientific discussions over Skype and our social experiences in both Charlottesville and Copenhagen.

Thanks to Susanne Kofoed, Jesper Mogensen, Andreas Klitgaard, Lisser St. Clair-Norton, Lone Hansen, Anna Joensen, Pernille Winther and Lou C. Svendsen for technical and administrative assistance during my PhD project. And thanks to Mikkel Lindegaard for assisting me with microarray experiments. I would also like to thank the people from the Sommer Lab for allowing use of laboratory equipment and especially thanks to Morten O. A. Sommer and Lejla Imamovic for setting up experiments concerning hypotheses on cancer cell metabolism. Also thanks to Mogens Kilstrup for many interesting discussions concerning bacterial metabolism. Thanks to Helle K. Johansen and Niels Høiby for allowing access to clinical strains (derived from the Copenhagen Universisity Hospital) and for helpful discussions about clinical data.

I would like to thank my friends and former PhD colleages: Nicholas Jochumsen, Søren Damkiær, Rasmus L. Marvig and Vinoth Wigneswaran for great company, scientific inspiration and support. And thanks also to the remaining members of the IMG and CfB groups, past and present: Heidi S. R. Johansen, Lei Yang, Yang Liu, Liang Yang, Anders Folkesson, Katherine Long, Martin W. Nielsen, Martin H. Rau, Sofia Feliziani, Fatima Y. Coronado, Rasmus Bojsen, Trine Markussen, Maria Gómez-Lozano, Eva K. Andresen, Claus Sternberg, Charlotte F. Michelsen, Christina I. A. Hierro, Anne-Mette Christensen, Hossein Khademi, Anders Norman, Sandra W. Thrane, Lea M. Sommer, Linda Jensen, Mikkel Lindegaard, Alicia Fernandez, Mette Munk and many more. Thank you for providing a great social working environment. It has been a pleasure working with you.

Last but not least, I would like to express my greatest gratitude to my family and friends for their endless love and support.

Abstract

Biological systems are complex. When we want to understand biological processes we often need advanced methods to reveal the relationship between genotype and phenotype.

The focus of this thesis has been to extract biological meaningful features from complex data sets and to use mathematical modeling to uncover how human pathogens adapt to the human host. *Pseudomonas aeruginosa* infections in cystic fibrosis patients are used as a model system for understanding these adaptation processes.

The exploratory systems biology approach facilitates identification of important phenotypes and metabolic pathways that are necessary or related to establishment of chronic infections. *Archetypal analysis* showed to be successful in extracting relevant phenotypes from global gene expression data. Furthermore, *genome-scale metabolic modeling* showed to be useful in connecting the genotype to phenotype at a systemic level. Particular metabolic subsystems were identified as important for metabolic adaptation in *P. aeruginosa*. One altered metabolic phenotype was connected to a genetic change; a finding that was possible through the systems characterization and which was not identified by classical molecular biology approaches where genes and reactions typically are investigated in a one to one relationship.

This thesis is an example of how mathematical approaches and modeling can facilitate new biological understanding and provide new surprising ideas to important biological processes.

Dansk resumé

Biologiske systemer er komplekse. Når vi ønsker at forstå biologiske processer, har vi ofte brug for avancerede metoder til at afsløre sammenhængen mellem en organismes arveanlæg og dens observérbare karaktertræk.

Fokus for denne Ph.d.-afhandling har været at ekstrahere biologiske meningsfulde egenskaber fra komplekse datasæt samt at bruge matematisk modellering til at afdække, hvordan sygdomsfremkaldende bakterier udvikler sig under kroniske infektioner i mennesket. Infektioner med bakterien *Pseudomonas aeruginosa* i cystisk fibrose patienter er anvendt som model-system til at forstå disse tilpasningsprocesser.

Den anvendte systembiologi-metode muliggør identifikation af vigtige karaktertræk og stofskiftereaktioner som er nødvendige eller relaterede til etablering af kroniske infektioner. *"Arketype-analyse"* viste sig at være nyttig til ekstrahering af relevante karaktertræk fra data. Derudover viste matematisk modellering af hele stofskiftet sig også at være nyttig til at forstå hvordan de overordnede stofskiftereaktioner i bakterien ændrer sig. Specifikke stofskiftereaktioner blev identificeret som værende vigtige for tilpasningen af *P. aeruginosa*. En specifik ændring i stofskiftet blev koblet til en genetisk ændring. Denne opdagelse var mulig ved hjælp af den anvendte systembiologiske metode, men havde derimod ikke været mulig gennem klassiske molekylærbiologiske metoder, hvor reaktioner og gener typisk undersøges i et en-til-en forhold.

Denne afhandling er et eksempel på, hvordan matematiske metoder og modellering kan føre til ny biologisk forståelse og bidrage med nye overraskende ideer til vigtige biologiske processer

Table of contents

List of publications

Thøgersen J C, Mørup M, Damkiær S, Molin S, Jelsbak L (2013). Archetypal Analysis of diverse *Pseudomonas aeruginosa* transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics*, **2013**, 14:279

Thøgersen J C*, Bartell J A*, Thykaer J, Nielsen K F, Johansen H K, Papin J A, Molin S, Jelsbak L (2015). Systems-based analysis of metabolic evolution during pathogen adaptation to the human host. *Manuscript submitted for publication. *Equal contribution.*

Not included in this thesis

Bartell J A, Blazier A, Yen P, **Thøgersen J C**, Jelsbak L, Papin J A. (2015). Reconstructing the metabolism metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. *Manuscript in preparation.*

Abbreviations

CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator
E. coli	*Escherichia coli*
FBA	Flux balance analysis
FVA	Flux variability analysis
H. influenzae	*Haemophilus influenzae*
P. aeruginosa	*Pseudomonas aeruginosa*
PCA	Principal component analysis
PCL	Periciliary liquid layer
PMN	Polymorphonuclear neutrophil
S. aureus	*Staphylococcus aureus*
SCFM	Synthetic cystic fibrosis sputum medium
SNP	Single nucleotide polymorphism

List of figures

Chapter 1

Introduction

Technological advances during the past 10-15 years have led to generation of numerous high-throughput data sets including whole-genome sequences and global gene expression data. The rapidly growing number of high-throughput data sets has raised a demand for new analytical tools that can assist in generating biological understanding from these data. There is a need for modeling approaches to assist in connecting genotype to phenotype at a systemic level and for analytical tools that can extract biological features from the complex high-throughput data sets, both of which are important elements of systems biology (Bordbar *et al*, 2014; Heinemann & Sauer, 2010).

Systems biology contains elements from chemistry, biology, engineering and computer science and it deals with integration of technology, biology and computation (Aebersold *et al*, 2000; Ideker *et al*, 2001). Systems biology can be described as the multidisciplinary approach to investigating the complexity of an organism (Hindré *et al*, 2012). In molecular biology genes are often investigated individually to assign a function to a single gene. Systems biology deals with the understanding of how molecular components collectively give rise to phenotype and physiology and in systems biology the interrelationship of all elements in a system is studied rather than studying them one at a time (Gunawardena, 2014; Hood, 2003).

One area where there is a need for connecting the genotype to a phenotype at a systemic level is in our pursuit of understanding pathogen behavior in terms of what makes a pathogen a pathogen and how does a pathogen evolve (de Lorenzo, 2015). A human pathogen is characterized by being able to colonize and grow within the human host and cause disease (Madigan & Martinko, 2006a). The latter is often due to virulence factors produced by the pathogen. There has been a lot of attention to virulence factors produced by human pathogens (Rahme *et al*, 1995; McDermott *et al*, 2011; Clatworthy *et al*, 2007; Eisenreich *et al*, 2010). However, it is also important to understand how the pathogens beyond virulence factors are able to grow within the human host and how they adapt to the host environment to ensure survival. Metabolism can be defined as all chemical reactions in a living organism including pathways necessary to degrade nutrients in the surrounding environment to obtain energy for survival and growth (Madigan & Martinko, 2006b). Therefore, in order to understand pathogen behavior in terms of how they survive in the host environment, it is necessary to

study metabolism and when we try to eradicate invading pathogens from the human host, insight into pathogen metabolism is crucial (de Lorenzo, 2014).

The opportunistic pathogenic bacterium *Pseudomonas aeruginosa* is an ideal model organism for understanding these processes. Environmental *P. aeruginosa* strains have optimized their metabolism for survival in their natural environment outside the human host. During long-term infections in the CF lung environment *P. aeruginosa* will most likely undergo adaptation to obtain a metabolic phenotype, which is optimal for survival in the human host environment. The identification of metabolic pathways that change during adaptation may therefore uncover, which metabolic pathways are important for pathogenesis. Since metabolism is a complex network of chemical reactions we need a systemic approach, which consider all reactions at ones. In general, identification of phenotypic patterns that characterizes pathogen persistence is desired.

I consider systems biology to fall within the field of research termed discovery science or exploratory research in contrast to hypothesis-driven research, but it can also be argued that it is a combination of both (Aebersold *et al*, 2000; Ideker *et al*, 2001). Hypothesis-driven research is well established and accepted and it is based on concrete hypothesis backed up by theory. Exploratory research on the other hand, examines unknown areas with little or no supporting theory and the fundamental idea is to explain variations observed in data without prior knowledge (Haufe, 2013; Waters, 2007). This can be a great advantage in providing new knowledge, but at the same time it is a big challenge for systems biology because the limited level of detail and mechanistic insight can be considered a weakness. Often the systems analysis is followed up by multiple hypothesis-driven experiments, which makes these studies quite comprehensive in terms of resources and collaborations across different academic disciplines. Therefore, when we model a system we often need to make compromises between the size of the system we wish to consider and the level of detail at which we model the system (Heinemann & Sauer, 2010).

Aim of project

The overall aim of this PhD project is to gain more knowledge in the adaptation process of human pathogens during chronic infections through a systems biology approach. The system biology approach consists in extracting biological knowledge from complex data sets through advanced exploratory methods. The project can be divided into two parts, which are represented by **Paper 1** (Chapter 5) and **Paper 2** (Chapter 6) respectively:

(1) The first project aims to identify patterns in global gene expression data sets through *Archetypal Analysis*, and to understand the biological meaning of these patterns related to the adaptation process of *Pseudomonas aeruginosa* during long-term airway infections of cystic fibrosis patients.

(2) The second project aims to identify which metabolic pathways that change during adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis lung environment through a combined experimental and computational approach including isotope-labeling experiments and genome-scale metabolic modeling.

Outline of thesis

The next chapter (Chapter 2) introduces *P. aeruginosa* and cystic fibrosis, which is followed by an introduction into feature extraction in data analysis in Chapter 3 and genome-scale metabolic modelling including the concept of isotope-labeling experiments in Chapter 4. The results of this PhD project are described in **Paper 1** (Chapter 5) and **Paper 2** (Chapter 6) and I recommend reading **Paper 1** and **Paper 2** before the discussion in Chapter 7.

Chapter 2

Pseudomonas aeruginosa and cystic fibrosis

Pseudomonas aeruginosa

We use the gram-negative bacterium *Pseudomonas aeruginosa* as a model organism to investigate pathogen adaptation in a human host environment. *P. aeruginosa* is an opportunistic pathogen, which means that it rarely infects healthy individuals, but it is a major cause of infections in patients with cystic fibrosis (CF) and it is also causing infections in immunocompromised individuals and people with severe burns and diabetes mellitus (Ramos, 2004). It is easy to isolate from the environment, where it is often found in soil, water and plants. Usually, the *P. aeruginosa* infections are acquired from the environmental sources, but transmission between CF patients also occurs (Jelsbak *et al*, 2007).

Cystic fibrosis

Cystic fibrosis is an autosomal recessive disorder caused by a mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. It is the most common inherited disease among Caucasians, where it appears in approximately one out of 2500 newborns (Folkesson *et al*, 2012). The CFTR gene encodes a chloride channel responsible for epithelial ion transport. Impaired function of this channel mostly affects the airways of the lungs, but it is also causing gastrointestinal, nutritional and other abnormalities (Lyczak *et al*, 2002; Gilligan, 1991). In healthy lungs, the airway epithelial cells are covered with a periciliary liquid layer (PCL) and an upper mucus layer (Figure 1). Together they form an essential part of the mucociliary clearance, which provides protection towards inhaled particles including microorganisms. Inhaled particles get stuck on the upper mucus layer and the PCL (separating the mucus layer from the epithelial cell surface) facilitates motility of cilia that serve as an escalator to remove the inhaled particles (Buchanan *et al*, 2009; Knowles & Boucher, 2002).

In CF patients, one consequence of the defect chloride channel is dehydration of the PCL and a thick mucus layer in the airways, which impairs the mucociliary clearance and the CF patients are therefore very susceptible to airway infections (Figure 1) (Lyczak *et al*, 2002; Gilligan, 1991; Govan & Deretic, 1996).

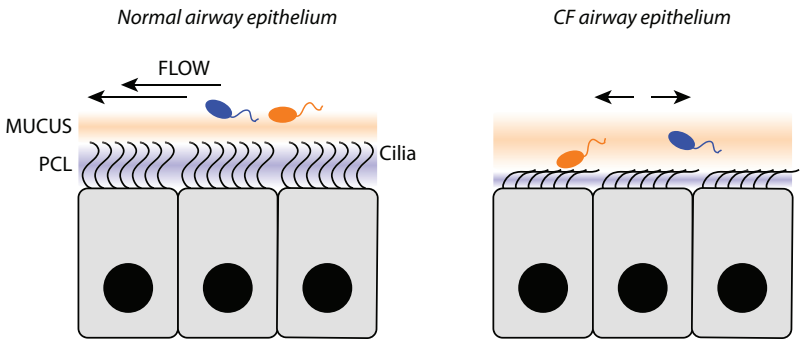


Figure 1. Comparison of the mucociliary clearance between normal airway epithelium and cystic fibrosis airway epithelium. **The normal airway epithelium is covered by a hydrated periciliary liquid layer (PCL) and on top of this thin mucus layer. Motion of the cilia ensures a unidirectional flow of mucus and thereby excretion of any inhaled bacterium. In the cystic fibrosis (CF) airway, a thick mucus layer covers the epithelium due to dehydration of the periciliary liquid layer. Motion of the cilia is impaired and bacteria can persist in the CF airway. The figure is modified from (Lyczak *et al*, 2002; Folkesson *et al*, 2012).**

Airway infections in cystic fibrosis patients

In addition to *P. aeruginosa*, a range of different microorganisms are associated with CF lung infections including the bacterial species *Staphylococcus aureus*, *Haemophilus influenzae*, *Streptococcus pneumonia*, *Burkholderia cepacia*, *Burkholderia pseudomallei* and also the fungal species *Aspergillus* and *Candida* are frequently observed (Burns et al, 1998; Foweraker, 2009). The prevalence of these species changes over time, where the first lung infections of the cystic fibrosis patients often appear in early childhood with *S. aureus* and *H. influenzae* and later, *P. aeruginosa* becomes most dominant (Harrison, 2007). Often the patients have recurrent acute infections with *P. aeruginosa*, but eventually these infections turn chronic where *P. aeruginosa* cannot be eradicated despite of host immune response and intensive antibiotic treatment. These chronic infections result in prolonged inflammatory response leading to destruction of the lung tissue and loss of lung function, and chronic infections with *P. aeruginosa* are the main cause of morbidity and mortality in CF patients (Döring et al, 2012).

Treatment and prognosis

The CF patients are treated with a range of different antibiotics. Eradication of early infection and prevention of chronic infection has been associated with clinical benefits (Döring *et al*, 2012). In industrialized countries, antibiotic therapy has played a major role in increasing mean life expectancy from 14 years for CF patients born in 1969 to more than 40 years for CF patients born in 2010 (Cystic fibrosis foundation patient registry 2009 annual data report, 2010; Döring *et al*, 2012). In the Copenhagen CF clinic the young CF patients with acute *P. aeruginosa* infections are treated occasionally

with antibiotics, whereas the patients chronically infected with *P. aeruginosa* continuously receive suppressive antibiotic therapy (Döring *et al*, 2000; Høiby *et al*, 2005).

The cystic fibrosis lung environment

When *P. aeruginosa* enters the airways of a CF patient it meets a lot of physical changes. The transition from the natural environment to the lung environment is characterized by changes in temperature, nutrient accessibility, gas composition (*e.g*. oxygen and carbon dioxide) and composition of surrounding microbes and presence of polymorphonuclear neutrophils (PMNs) of the immune system (Hauser *et al*, 2011). In addition to that, *P. aeruginosa* is exposed to a range of antibiotics while being inside the human airways as described above. It has been shown that *P. aeruginosa* grows within the characteristic CF sticky mucus, where it meets hypoxic conditions (Ohman & Chakrabarty, 1982; Worlitzsch *et al*, 2002; Palmer *et al*, 2005). The nutrient composition of the CF mucus has previously been analyzed and a synthetic CF sputum medium (SCFM) was defined to mimic the CF mucus composition (Palmer *et al*, 2007). The CF mucus is characterized by being rich in amino acids and it also contains glucose and lactate (Palmer *et al*, 2007). Although being a nutrient-rich environment, the CF lungs also provide stressful conditions for invading pathogens due to the frequent antibiotic treatment and host immune response (Folkesson *et al*, 2012). Oxidative stress results from reactive oxygen species (ROS) produced by PMNs as part of the host immune response to infection (Hull *et al*, 1997; Høiby, 2006). Most recently, it has also been suggested that antibiotics may induce intercellular ROS production in bacteria (Dwyer *et al*, 2014; Kohanski *et al*, 2010). The presence of oxidative stress can induce bacterial mutation rates and select for variants that are less sensitive to oxidative stress (Folkesson *et al*, 2012).

Adaptation of Pseudomonas aeruginosa

The ability of *P. aeruginosa* to thrive and persist in the human lung environment is facilitated through its large metabolic versatility and regulatory genes (Ramos, 2004), but also through genetic mutations as observed for many clinical isolates of *P. aeruginosa* (Smith *et al*, 2006; Yang *et al*, 2011; Dettman *et al*, 2013).

The adaptation process of *P. aeruginosa* is interesting from several points of view. First of all, knowledge on adaptation can guide new therapeutic intervention. For example reactions that become essential for adapted strains of *P. aeruginosa* compared to the original infecting strains can serve as targets for future antibiotics. Second, the fact that *P. aeruginosa* persists in the lung environment over many decades allow us to study long-term evolution of a pathogen inside the human

host. The knowledge gained from adaptation studies of *P. aeruginosa* is also relevant for industrial applications. The process by which *P.* aeruginosa reshapes its metabolism to fit the new environment can be parallel to the way we wish to engineer a production organism to obtain a certain phenotype in an industrial setting.

Past studies of within-host evolution of Pseudomonas aeruginosa

The transition from initial colonization to chronic infection of *P. aeruginosa* is often followed by observed phenotypic changes in *P. aeruginosa* caused by genetic adaptation to the CF lung environment. Frequently observed phenotypic changes include a slow-growth phenotype (at least *in vitro*), conversion to a mucoid phenotype, gain of antibiotic resistance, loss of motility, loss of quorum sensing, appearance of small colony variants, increased mutation rate (hereafter 'hypermutators'), decreased production of virulence factors and cell envelope changes (Harrison, 2007; Burns *et al*, 1998; Lyczak *et al*, 2002). Among the listed phenotypes are the mucoid phenotype and the hypermutator that I will return to in the discussion.

The mucoid phenotype is most often caused by a mutation in the anti-sigma factor *mucA* and it is characterized by a high production of alginate, which is easy to detect on laboratory growth plates, since the bacterial colonies appear very moist and sticky (Ciofu *et al*, 2008; Rau *et al*, 2010). Alginate is thought to protect the bacteria from phagocytosis by neutrophils and macrophages and to resist oxidative stress (Mathee *et al*, 1999; Oliver & Weir, 1985). However, a conversion back to a non-mucoid phenotype also appears in clinical isolates, why it is speculated that it may not be the alginate production itself providing the adaptive advantage, but maybe alginate production is a secondary effect of the *mucA* mutation (Damkiær *et al*, 2013; Rau *et al*, 2010).

Hypermutators are frequently observed among isolates of *P. aeruginosa* isolated from chronic CF infections, but the hypermutators are also widespread among other pathogenic bacteria suggesting that being a hypermutator is an advantage in pathogenesis (Weigand & Sundin, 2012; Oliver & Mena, 2010; Hogardt *et al*, 2007; Sundin & Weigand, 2007). The mutator phenotype has been associated with an evolutionary advantage during bacterial adaptation to new environments or stressful conditions (Oliver & Mena, 2010). Hypermutation creates genetic and phenotypic variations in a population and it is suggested that hypermutation is a mechanism for acceleration of bacterial evolution (Oliver & Mena, 2010).

Over the past years, there has been an increasing number of genomic studies of *P. aeruginosa* with the aim of understanding pathogen behavior inside the human host (Tümmler *et al*, 2014; Yang *et al*,

2011; Marvig *et al*, 2015). One extensive study from 2011 (Yang *et al*, 2011) (including genome sequencing, transcriptional profiling and phenotypic arrays) characterizes the within-host evolution of a particular *P. aeruginosa* clone type, *P. aeruginosa* DK2 (formerly known as the 'b' clone), which has successfully been transmitted between patients attending the Copenhagen CF clinic (Jelsbak *et al*, 2007). The DK2 clone was isolated for the first time in the Copenhagen CF clinic in 1973. Since then it has caused chronic infections in several patients and it has now persisted in the CF lung environment for four decades. The study of multiple DK2 strains isolated from 1973 to 2007, revealed that the DK2 lineage underwent an initial period of rapid adaptation before 1979 and important mutations were identified in global regulatory genes (Yang *et al*, 2011). Phenotypic characterization of catabolic performances showed that the adapted phenotype had lost its catabolic function on various carbon and nitrogen sources compared to the phenotype of the initial strains (Yang *et al*, 2011).

In a recent study a list of 52 so-called pathoadaptive genes were identified among 474 genome-sequenced clinical *P. aeruginosa* isolates representing 36 different lineages (Marvig *et al*, 2015). Pathoadaptive genes are described as genes in which mutations optimize pathogen fitness and they are identified as genes in which mutations appear very frequently in the 36 lineages. The list of pathoadaptive genes includes genes that are well-known to be involved in adaptation. The list also includes genes potentially important for adaptation, which have not been emphasized previously. This type of study is very valuable in creating an overview of common genetic adaptation. However, we must not neglect mutations in genes that may not appear often enough to be considered among the pathoadaptive genes, since diverse genetic mutations could result in the same functional effect. We therefore need also to focus our attention to adaptive changes that appear on a functional level - for example on metabolism. However, little is known about metabolic adaptation of *P. aeruginosa* inside the CF lung and in general *in vivo* metabolism of bacterial pathogens is poorly understood (Eisenreich *et al*, 2010).

A well-annotated genome will give us information about the organism's metabolic repertoire in terms of a list of metabolic reactions that are possible based on proteins encoded in the genome. A study of metabolism will add information on pathway activities and understanding of which pathways are essential for survival under given conditions. Identification of pathways necessary for within-host pathogen survival will bring us a step closer to new treatment strategies.

Present study: applying systems biology tools in the data interpretation process

Traditional long-term evolution experiments carried out *in vitro* focus on adaptive events connected to defined selective pressures (Cooper *et al*, 2003; Elena & Lenski, 2003). For our model system

many unknown parameters may influence the adaptation and it is therefore a more complex task both to identify the adaptive phenotypes and to define the relevant selective pressures present in the CF lung environment.

When *P. aeruginosa* adapts to the CF lung environment it will adapt in response to many changed factors at one time. It is far from the simple case where adaptive evolution is observed in response to for example a single antibiotic or carbon source in a test tube experiment. Therefore the analysis of *P. aeruginosa* adaptation is also complicated. We need to consider many objectives and here the exploratory analysis as discussed in Chapter 1 may have its benefits. Instead of formulating clear hypotheses about adaptation, we ask what we can learn about adaptation from the available data sets.

As described above, identification of pathoadaptive genes can be done without focusing directly on the function of the genes. Instead multiple adapted strains are compared and if the same single nucleotide polymorphism (SNP) appears more frequently than it would be expected statistically it is indicating that the SNP has a function for adaptation (Feliziani *et al*, 2014; Marvig *et al*, 2015). Although these individual SNPs may be linked to phenotypes through genome annotations, we will not be able to deduce how the combination of SNPs may affect the phenotype of the organism. Experimentally, it is possible to genetically engineer a SNP or a combination of SNPs into a wild type strain and compare the physiology between the strain with and without the mutation. However, one challenge connected to studying metabolism is that metabolism changes rapidly in response to environmental changes. Therefore, it can be difficult to measure the relevant metabolic phenotype, since the growth environment can very likely influence the effect of the SNP, and the conditions in the laboratory may not be optimal to study an effect within the human airways.

Our ultimate aim is to understand how metabolism changes during adaptation in the CF lung environment at a systemic level. One way to accomplish that can be to model this effect *in silico* by integrating information from genetic mutations and gene regulation through global gene expression studies into a genome-scale metabolic model of *P. aeruginosa*.

The next chapters will introduce different analytical methods that can be used to extract information from complex high-throughput data sets and introduce the concepts in genome-scale metabolic modeling.

Chapter 3

Data analysis - Feature extraction from complex data sets

High-dimensional datasets are essential to systems biology. When you deal with high-dimensional data sets it is often desired to reduce the number of dimensions in order to extract features of the data set. The general principle is to identify a few components that can represent the data set and each point in the data set can then be described by a combination of these components.

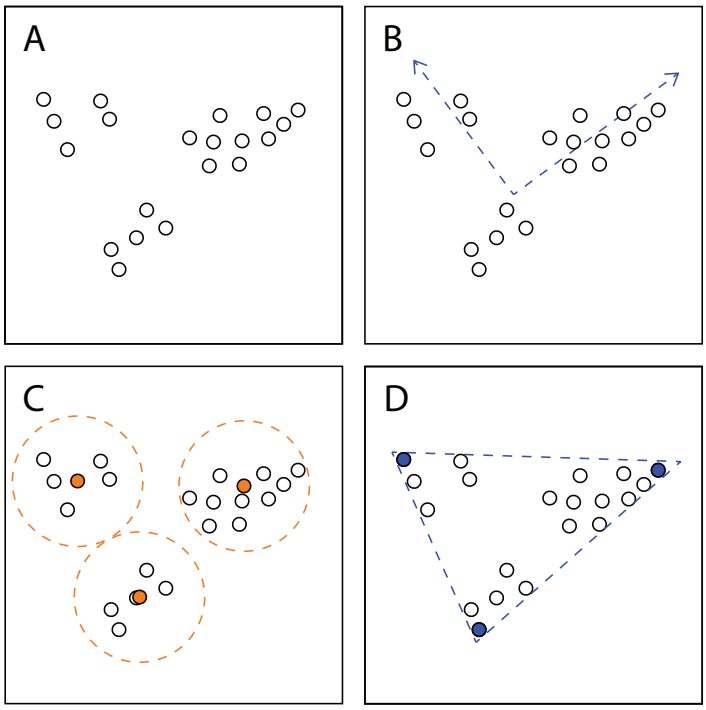


Figure 2. Illustration of dimension-reduction techniques.

A. Raw data. The original data set is composed of 20 data points.

B Principal component analysis (PCA). The principal components are defined as orthogonal vectors in the orientation of maximum variance.

C K-means clustering. Three clusters are represented by cluster centroids (orange dots), which are the mean values of the data points within each cluster.

D Archetypal analysis. Three archetypes are representing the data set as extreme points (blue dots). All data points can be described as a combination of archetypes.

Imagine a case where we have 20 data points. It would be much easier if we could reduce the number of points that we have to interpret. If we can describe the whole data sets by finding for example three representative points (or components) we have succeeded in making a dimension reduction, which facilitates easier interpretation.

This chapter will provide some conceptual background on the commonly used dimension-reduction techniques, principal component analysis (PCA) and k-means clustering, and also introduce the concept of *archetypal analysis*, which is applied for gene expression data analysis in **Paper 1**. For simplicity, I have decided to leave out mathematical formulas. However, they appear in the Materials

and Methods section of **Paper 1**. The concepts of PCA, k-means clustering and archetypal analysis are illustrated in Figure 2.

Principal component analysis

One widely used dimension-reduction technique is principle component analysis. In PCA, the components are defined in such a way that they represent most variance present in the dataset, while being orthogonal to each other. The data is transformed into a new coordinate system, where the first axis is placed in direction of most variance and this is defined as the first principal component (PC1). The second component (PC2) is orthogonal to PC1 and again oriented to account for most variance given the orthogonality constraint to PC1. The number of components equals the number of data points and the subsequent components are defined similar to PC1 and PC2. Each data point will be assigned a coordinate for each principal component. The data can then be visualized by plotting data as a combination of two or three principal components. This visualization possibility of otherwise complex data sets makes PCA popular (Friedman *et al*, 2009). PCA has a large degree of flexibility and it is perfect for capturing variance in data. However, the principal components may be hard to characterize due to their complex representation (Mørup & Hansen, 2012).

K-means clustring

K-means clustering is a method that separates all observations in a data set into a pre-defined number of subsets or clusters. For each defined cluster the mean of observations is defined as the cluster centroid (or component). The k-means clustering algorithm identifies the clusters, so that each data point within a cluster is closest to its own cluster centroid (Friedman *et al*, 2009). The benefit of clustering approaches is that the features of the cluster components are similar to the data and this makes the results easier to interpret. However, the rigid assignment of data points into a single cluster may cause loss of information connected to similarities between data points assigned to different clusters (Mørup & Hansen, 2012).

Archetypal analysis

Archetypal analysis estimates the principle convex hull of the data set, which can be described as a minimal set of points that can wrap a given data set (Cutler & Breiman, 1994). Archetypal analysis is different from k-means clustering in that it identifies the extreme (but representative) data points (archetypes) in the data. Each data point can then be described as a combination of archetypes and each archetype can be described as a combination of data points. Archetypal analysis thereby also

includes the features of factorization like PCA. In archetypal analysis a pre-defined number of components are identified, so that each data point is best described by a combination of these components (Friedman *et al*, 2009). The archetypes will often be placed at the surface of the convex hull and the advantage of archetypal analysis is that these extreme profiles are more likely easier to characterize (Mørup & Hansen, 2012).

PCA, k-means clustering and archetypal analysis all fall within the statistical category unsupervised learning. The goal of unsupervised learning is to reduce dimensionality of data, cluster data and to find the hidden sources and causes of the data (Roweis & Ghahramani, 1999). In **Paper 1** we apply these three methods for analysis of gene expression data. We use the methods to identify patterns or similarities between samples (each of which has its own gene expression profile). Thereby we find expression patterns that represent the total data set instead of analyzing each expression profile individually. The use of different unsupervised learning techniques can be compared to visualizing a problem wearing different sets of glasses or from different angles.

Chapter 4

Genome-scale metabolic modeling

The increasing amount of genome-scale data and whole genome sequences since the 1990s have enabled development of genome-scale metabolic models accounting for metabolism at a systemic level (Covert *et al*, 2001). Genome-scale metabolic models are mathematical representations of the interactions between metabolites, enzymes, and genes that enable metabolic activity such as production of biomass and synthesis of important byproducts through the catabolism of growth substrates in the environment.

The first published genome scale metabolic model was developed for *H. influenzae* and published in 1999 by Edwards and Palsson (Edwards & Palsson, 1999). Since then various models have been developed for different organisms and updated versions of the models are generated as new information becomes available (Feist *et al*, 2009). The first genome scale metabolic model of *P. aeruginosa* was published in 2008 (Oberhardt *et al*, 2008). This chapter will introduce the concepts of genome-scale metabolic modeling and round off with application examples.

Genome-scale metabolic model reconstruction

The process of reconstructing a genome scale metabolic model is visualized in Figure 3. First, a complete list of reactions present in the organism of choice is defined based on the genome annotation and through review of databases and literature (Oberhardt *et al*, 2009). The model can be built manually or a draft model can also be generated through different semi-automated tools *e.g.* the online generation of genome-scale metabolic models using the tool *'Model SEED'* (Overbeek *et al*, 2005). The quality of the first model draft will rely on how well the genome annotation is and how much is known about metabolism in the organism from the literature. The modeling process will also give rise to improved annotation through identification of incomplete pathways for example connected to catabolism of substrates (known to be degraded in the organism) or production of metabolites that are observed experimentally (Bartell *et al*, 2014; Monk *et al*, 2013). The resulting genome-scale model can be considered as a collection of all known metabolic pathways in the organism including information about which enzymes are catalyzing the reactions linked to genes that are encoding the enzymes - the so-called Gene-Protein-Reaction (GPR) relationship. The physical product of the model at this stage could be an excel spreadsheet with reactions listed in rows and for each

reaction information about metabolite substrates and products and genes related to that reaction. The next step is to convert the model into a stoichiometric matrix $\underline{S}$, which is the mathematical representation of the model. The stoichiometric matrix $\underline{S}$ is an *m x n* matrix where *m* equals the number of metabolites and *n* equals the number of reactions. The numbers inside the $\underline{S}$ matrix indicates the stoichiometry between substrates and products for a given reaction. For all metabolites that are not involved in the given reaction the values in $\underline{S}$ will be listed as zeros, substrates will take negative values and products will take positive values (Figure 3B) (Becker & Palsson, 2008; Oberhardt *et al*, 2009).

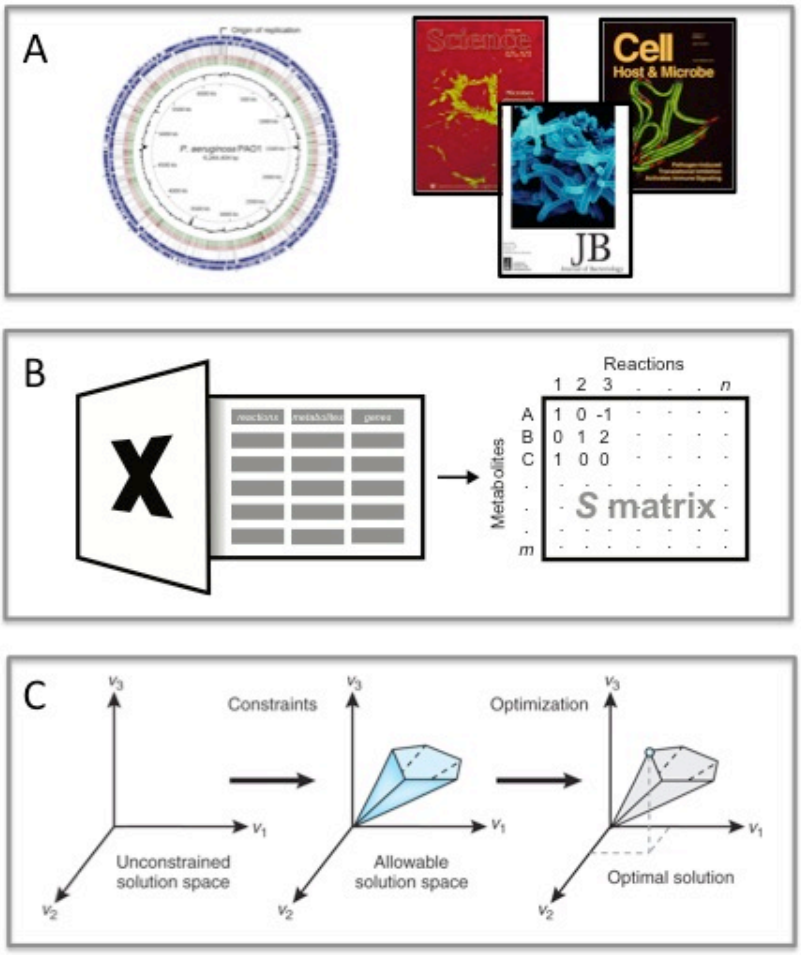


Figure 3. Genome-scale metabolic modeling.
A. Genome annotations and information from literature form the base of the metabolic model. The genome map is from (Weigand & Sundin, 2012).
B. All metabolic reactions are listed in Excel including connections between reactions, metabolites and genes. This information can be translated into a stoichiometric matrix (S), which is the mathematical representation of the model. **C.** Fluxes can now be calculated and constraints can be applied to the model to find the optimal solution. The figure in panel C is modified from (Orth *et al*, 2010).

In silico flux predictions

Flux balance analysis (FBA) is an *in silico* prediction of flux profiles through a metabolic network. FBA is based on the assumption that all metabolites are at steady state. That means that the metabolite concentrations are constant, and hence that the sum of reaction fluxes producing or importing a given metabolite equal the sum of reaction fluxes that are consuming or exporting the same metabolite. Mathematically it can be written as Equation 1, where $\underline{\underline{S}}$ is the Stoichiometric matrix and $\underline{v}$ is the flux vector, which represents the fluxes through all reactions in the network (Varma & Palsson, 1994; Edwards & Covert, 2002).

$$\underline{\underline{S}} \cdot \underline{v} = \underline{0} \qquad \text{(Equation 1)}$$

$$a_i < v_i < b_i \qquad \text{(Equation 2)}$$

$$\text{Max}(v_{\text{biomass}}) \qquad \text{(Equation 3)}$$

Experimentally, the steady-state requirement can be achieved under exponential growth. FBA uses linear programming to determine a solution of fluxes through the metabolic network that fulfill this steady state assumption and also additional defined constraints while optimizing an objective function (for example maximizing biomass production) (Haggart *et al*, 2011).

Typical constraints applied to FBA are defined reaction boundaries (Equation 2). If unconstrained, these can be mathematically defined as lower and upper bounds (a_i and b_i respectively) being arbitrarily set to for example -1000 and +1000. Some bounds can be constrained based on literature and experiments. Knowledge about irreversible reactions can for example change the lower bounds to zero. Both upper and lower bounds can also be set to zero to simulate a gene (or reaction) knockout.

The most commonly used objective function in constrained based modeling like FBA is maximizing biomass production (the rate at which metabolic compounds are converted into biomass components such as nucleic acids, proteins and lipids) (Orth *et al*, 2010). The biomass objective finds the flux solution with the highest *in silico* biomass yield among the possible solutions to the network (Equation 4). However, other objective functions can be defined *e.g.* minimizing or maximizing ATP production, and even a combination of objective functions can be defined (Schuetz *et al*, 2007).

When evaluating fluxes through a system it can also be relevant to perform a flux variability analysis (FVA). FVA determines the variability of each reaction, while maintaining the optimal objective function value (e.g. maximum biomass production level). The maximum objective function value is first determined through FBA, which allows reaction fluxes within the defined bounds. Then, for each

reaction the minimum and maximum reaction fluxes are determined with the constraint applied, that the maximum objective function value should be maintained. FVA thereby gives you the range of possible solutions for a given reaction to fulfill the objective function. The study of FVA will give a more robust analysis of metabolic activity since specific reaction fluxes proposed by the model through FBA could be one of many solutions to the objective function, whereas the FVA is more the solution space (Haggart *et al*, 2011).

Integration of data sets into genome-scale metabolic models

The genome-scale models can be combined with transcriptomic data or other omics data, which will add an additional layer of constraints for reaction fluxes (Bordbar *et al*, 2014). The model only accounts for metabolic genes; that means genes that are directly involved in the metabolic reactions. However, integration of gene expression data will make a fingerprint of regulatory events causing reactions to be inactive due or active due to down-regulated or up-regulated gene expression.

A variety of transcriptome integration tools have been developed over the past 10 years (Becker & Palsson, 2008; Zur *et al*, 2010; Jensen *et al*, 2011; Machado & Herrgård, 2014). Among the tools are both discrete and continuous integrations, but there is no clear preference when the methods are compared (Machado & Herrgård, 2014). One option is to develop proposed 'off' and 'on' gene activity levels based on gene expression values. Although gene expression levels do not necessarily reflect flux levels, the gene expression values can help to guide the determination of the correct phenotype among the space of solutions from the metabolic network (Machado & Herrgård, 2014).

Another layer of constraints can be based on genetic variations. If we wish to model differences in metabolic phenotypes due to genetic variations between strains, we need to define how genetic variations should be implemented into the models. When we look at SNPs in light of genetic adaptation of an organism there are different ways of evaluating whether the SNP has any functional effect on the organism. The SNPs can be divided into silent, missense and nonsense mutations. The silent mutations do not result in an amino acid change in the encoded protein and it is "tempting" to interpret this, as the SNP has no functional effect. However, silent mutations may have an impact on regulation, why they shouldn't just be neglected. For example a silent mutation in *Mycobacterium tuberculosis* has been associated with antibiotic resistance by converting a region adjacent to the silent mutation into an alternate promoter (Ando *et al*, 2014). Missense mutations on the other hand do result in a change in amino acid composition in the encoded protein. In order to evaluate the effect of a missense mutation, it can be relevant to look at the properties of the original and substituting amino acid. In 2009 an algorithm called SIFT *("Sorting Intolerant from Tolerant"*) was

developed with the aim of predicting effects of coding missense SNPs on protein function (Kumar *et al*, 2009). The algorithm takes the amino acid properties into consideration when calculating a score for the probability of the protein function being affected. Finally, nonsense mutations result in frame-shifts of the mRNA translation and the resulting amino acid sequence after a nonsense mutation must be assumed dysfunctional. The impact of a nonsense SNP therefore depends on where the SNP is located in the sequence, where an early nonsense SNP most likely result in a non-functional protein, whereas a late nonsense SNP can result in impaired or sustained function of the protein.

For our study presented in **Paper 2** we have chosen to define three different constraints that should resemble SNP impact. All of the reaction bound constraints are based on original FVA ranges calculated without implementing constraints due to SNPs. Then, these FVA ranges are adjusted in a way to resemble the likelihood of functional impact of the SNP. Thus, silent SNPs and missense SNPs that are predicted to be tolerated by the SIFT algorithm are implemented as constraints for the affected reactions defined as 10% reduced FVA range. Missense SNPs that are predicted to affect protein function are implemented with constraints defined as 50% reduced FVA range. Finally, nonsense mutations are implemented as a constraint reducing FVA range of the affected reaction with 90%.

Constraints can also be stated based on known fluxes through the network. Isotope labeling substrates can be used to determine fluxes through different convergent pathways (del Castillo *et al*, 2007; Nanchen *et al*, 2007; Wiechert, 2001). The concept is illustrated in Figure 4 and is briefly described below.

When a heterogeneous labeled substrate as [1-^{13}C]-glucose (glucose labeled with the ^{13}C carbon isotope at position 1) is catabolized, different labeling patterns on downstream intermediates (*e.g.* pyruvate) will occur depending on which pathway glucose has been degraded through (Figure 4). The degradation of glucose into pyruvate can happen through three alternate pathways; The *Embden-Meyerhof Parnas* (EMP) pathway (standard glycolysis in textbooks), the *pentose phosphate* (PP) pathway and finally the *Entner Doudoroff* (ED) pathway. If one glucose molecule is degraded through the EMP pathway the labeled carbon will end up in the third position for one out of two pyruvate molecules. If one glucose molecule is degraded through the PP pathway, the labeled carbon atom will end up in CO_2 and no labeled pyruvate will occur. If one glucose molecule is degraded through the ED pathway the labeled carbon will end up in the first position for one out of two pyruvate molecules. The resulting mix of pyruvate molecules with different labeling patterns can therefore be applied to calculate how much glucose is degraded through each of the pathways (Thøgersen, 2010). If we know the relative activities of these pathways, this information can be integrated into the

model by constraining the reactions to attain these values. Labeling patterns on other intracellular metabolites will likewise be useful for determining reaction fluxes.

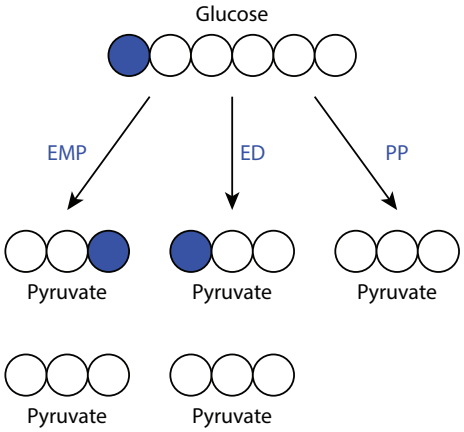


Figure 4. Isotope-labeling experiment. Glucose can be degraded into two molecules of pyruvate through three alternate pathways: Embden-Meyerhof-Parnas pathway (EMP), Entner-Doudoroff pathway (ED) and Pentose Phosphate pathway (PP). The labeled carbon atom from [1-^{13}C]-labeled glucose ends up at different positions in pyruvate dependent on which pathway is used to degrade glucose. Figure adapted from (Thøgersen, 2010).

Applications of genome-scale metabolic models

One major contribution from genome scale metabolic models is derived from the model reconstruction process itself. The reconstruction of the model will give rise to improved genome annotation. This happens through identification of pathways that are known to operate in the organism, but which are not complete in the *in silico* model. Searching for genes encoding proteins that are known to catalyze the relevant reactions can complete those pathways. Through directed blast analysis new gene annotations can thus be made. The manual curation process of the model will make the researcher relate information from the literature to databases and sometimes clarify discrepancies (Oberhardt *et al*, 2008; Monk *et al*, 2013).

The application of the genome-scale metabolic models that is closest to our use in **Paper 2** is contextualizing high-throughput data sets, where the model is used as a tool for analyzing high-throughput data and for predicting the systemic impact of information stored in the data. The model as a tool thereby allows us to interpret these high-throughput data at a network level. One example where genome-scale modeling is used to contextualize data is in the study by Lobel et al (2012), where

transcriptome data is combined with genome-scale metabolic models to define metabolic requirements of *Listeria monocytogenes* during infection (Lobel *et al*, 2012). Another study integrate gene expression data to make comparisons between *P. aeruginosa* strains representing different stages of adaptation and they make *in silico* predictions to determine production of virulence factors, growth capabilities and essential genes (Oberhardt *et al*, 2010).

The models can also be used for comparative modeling analysis where two or more models are compared to evaluate for example differences in metabolic capacity between models. One example is the study by Bartell et al (Bartell *et al*, 2014) where they compare metabolic capabilities and contextualizes genetic differences between *Burkholderia cenocepacia* and *Burkholderia multivorans*. Another study compare 55 genome-scale metabolic models of *Escherichia coli* in order to characterize the pan (collective) and core (shared) metabolic capabilities of the *E. coli* species (Monk *et al*, 2013). Through their comparative analyses, they are able to group the individual strains into correct pathotypes and environmental niches based on the *in silico* strain-specific metabolic capabilities and they suggest that a small number of nutrient sources can be used to classify different *E. coli* types (Monk *et al*, 2013).

Genome-scale metabolic models are often used in the biotech industry for making predictions for optimized product yield or quality that is useful prior to metabolic engineering (Puchałka *et al*, 2008; Kjeldsen & Nielsen, 2009). One example is to apply the model for making predictions about reduced byproduct formation. It is possible that knockout of certain genes will disrupt the byproduct formation. The models allow the researcher to make *in silico* knockout studies, which is less resource demanding than the corresponding *in vitro* experiments of gene knockouts. The model can help to select among a list of candidate genes based on the model predictions and the number of experimental knockouts can therefore be reduced. Another application of genome scale modeling in an industrial context could be identification of missing pathways for production of a desired product. The missing pathways could be constructed *in silico* and the model analysis could indicate whether the reactions are feasible in the context of the available metabolic pathways in the organism. In the study by Kjeldsen and Nielsen, they used genome-scale metabolic modeling for optimizing lysine yield in the bacterium *Corynebacterium glutamicum* (Kjeldsen & Nielsen, 2009).

The model can also be used to make *in silico* predictions of essential genes, which is highly relevant when studying pathogens, since essential genes can serve as targets for therapeutic intervention (Oberhardt *et al*, 2008; Bartell *et al*, 2014; Wodke *et al*, 2013). *In silico* essential gene analysis is carried out by removing one gene at a time from the model. If the model fails to produce biomass after

the deletion, the gene is called *in silico* essential. Some times more genes account for one reaction and this reaction will only be disrupted if all of the genes (with an 'or' relationship) are removed from the model. Therefore it is also interesting to perform an essential reaction analysis, where one reaction at a time is removed from the model in order to identify *in silico* essential reactions. The analysis of essential genes is often used to validate genome scale metabolic models since these values are easy to compare with experimentally determined essential genes (Oberhardt *et al*, 2008; Monk *et al*, 2013).

Genome-scale metabolic modeling is applied in **Paper 2** to contextualize the impact of SNPs and altered gene expression on metabolism at a systemic level. Subsystems subject to metabolic changes are identified as well as genes that become essential during the adaptation of *P. aeruginosa* to the CF lung environment. The modeling approach is also utilized to make predictions about the selective forces within the CF lung environment and to evaluate the feasibility of suspected metabolic changes.

Chapter 5

Paper 1

Archetypal Analysis of diverse *Pseudomonas aeruginosa* transcriptomes reveals adaptation in cystic fibrosis airways

Thøgersen J. C., Mørup M., Damkiær S., Molin S., Jelsbak L.

BMC Bioinformatics, **2013**, 14:279

Thøgersen et al. BMC Bioinformatics 2013, 14:279
http://www.biomedcentral.com/1471-2105/14/279

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Archetypal analysis of diverse *Pseudomonas aeruginosa* transcriptomes reveals adaptation in cystic fibrosis airways

Juliane Charlotte Thøgersen[1], Morten Mørup[2], Søren Damkiær[1], Søren Molin[1] and Lars Jelsbak[1*]

Abstract

Background: Analysis of global gene expression by DNA microarrays is widely used in experimental molecular biology. However, the complexity of such high-dimensional data sets makes it difficult to fully understand the underlying biological features present in the data.
The aim of this study is to introduce a method for DNA microarray analysis that provides an intuitive interpretation of data through dimension reduction and pattern recognition. We present the first "Archetypal Analysis" of global gene expression. The analysis is based on microarray data from five integrated studies of *Pseudomonas aeruginosa* isolated from the airways of cystic fibrosis patients.

Results: Our analysis clustered samples into distinct groups with comprehensible characteristics since the archetypes representing the individual groups are closely related to samples present in the data set. Significant changes in gene expression between different groups identified adaptive changes of the bacteria residing in the cystic fibrosis lung. The analysis suggests a similar gene expression pattern between isolates with a high mutation rate (hypermutators) despite accumulation of different mutations for these isolates. This suggests positive selection in the cystic fibrosis lung environment, and changes in gene expression for these isolates are therefore most likely related to adaptation of the bacteria.

Conclusions: Archetypal analysis succeeded in identifying adaptive changes of *P. aeruginosa*. The combination of clustering and matrix factorization made it possible to reveal minor similarities among different groups of data, which other analytical methods failed to identify. We suggest that this analysis could be used to supplement current methods used to analyze DNA microarray data.

Keywords: Archetypal analysis, Gene expression, *Pseudomonas aeruginosa*, Cystic fibrosis, Hypermutators

Background

DNA microarrays simultaneously monitor expression levels of thousands of genes, and this technology is widely used in experimental molecular biology. However, the complexity of such high-dimensional data sets makes it difficult to fully comprehend the underlying biological features present in the data. Different dimension reduction techniques aim to find patterns in high complexity data sets. The choice of analytical method can influence the interpretation of the data, and it can be useful to combine different methods.

K-means clustering and principal component analysis (PCA) are techniques for unsupervised pattern recognition commonly used in microarray data analysis. K-means clustering aims to group samples (or genes) with similar behavior [1]. Each sample is then assigned to a cluster represented by a cluster centroid. PCA is an orthogonal linear transformation transforming the data into a new coordinate system where the axes are oriented to account for maximal variation in the data set. PCA decomposes data into a set of uncorrelated variables called principal components [2-4].

Clustering approaches give easy interpretable features but pay a price in terms of modeling flexibility, because each sample must be grouped in only one cluster and no intermediate between clusters is allowed. PCA on the

* Correspondence: LJ@bio.dtu.dk
[1]Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark
Full list of author information is available at the end of the article

other hand can lead to complex representations from which we learn relatively little about the data. Archetypal analysis (AA) combines the virtues of both clustering and PCA in that AA results in easy interpretable components that have added flexibility over clustering by allowing intermediates [5]. Cutler and Breiman first introduced AA in 1994, where they used AA to analyze air pollution and head shape [6]. Later, AA has been applied in the identification of extreme practices in benchmarking and market research and signal enhancement and feature extraction of IR image sequences [7,8]. Recently, AA has been shown to be useful in extracting features from different high-dimensional data sets including neuroimaging, computer vision and text mining data sets [5] and also in identifying extreme and representative human genotypes within the human population [9].

AA estimates the principle convex hull of a data set. The convex hull can be described as a minimal set of points that can wrap a given data set. The idea of AA is to find a few representative points (archetypes) in a data set such that all data can be described as a convex combination of these archetypes. The archetypes are related to experimental data but they are not necessarily observed points in the data set. Each archetype represents distinct characteristic features. Explaining data as a combination of these features can make the data set easier to interpret [5]. Unlike PCA, AA is not restricted by orthogonality, and it is possible that this method will clarify biologically meaningful features that are not discovered by PCA, while resulting in a more detailed account of the data than given by clustering approaches such as k-means clustering.

AA has been shown to be useful in extracting features from different high-dimensional data sets. So far, the method has not been applied to gene expression data despite clear advantages such as the intuitive and straightforward interpretation of the AA components. AA can be considered an unmixing approach that decomposes each observation into a weighted average of features defining distinct aspects in the data. In the related unmixing framework for gene expression data proposed in [10] the data is projected to a PCA subspace. In this subspace each observation is defined as convex combinations of features forming the simplex with smallest volume among candidate simplices that are found by an iterative boundary growing procedure that is terminated when all observations are enclosed. Contrary to this framework, AA operates directly on the full data and as the features are constrained to be convex combinations of the observations the archetypes will not in general enclose all observations.

Variation of phenotypes found in nature has recently been described as weighted averages of archetypes, where archetypes represent phenotypes that are optimized for a single adaptive task [11]. The phenotype space will often be arranged in a simple geometric shape where archetypes represent the corners, and the closer a point is to a corner the more important the corresponding task is to fitness in the organism's habitat [11]. From this it can be concluded that it is possible to identify the tasks that are important for fitness by analyzing these corners [12]. Furthermore, the variation within a species (the combination of archetypes) reflects the different environments it inhabits [11]. The message of the paper by Shoval et al. (Science) [11] clearly illustrates the value of AA and the idea of considering a phenotype space as a combination of extreme but representative points, which is exactly the concept of this present analysis: Archetypal Analysis.

In this study, we apply AA to five gene expression data sets for *Pseudomonas aeruginosa* isolated from the lungs of cystic fibrosis patients. The five data sets were based on different experimental conditions including growth medium and growth state during cell harvesting. A method like PCA most likely captures this experimental variance in the first few components. The first components will make restrictions for the additional components due to the orthogonality constraint, and information that is linked to the real biological difference between the samples may be difficult to extract. Since AA is not restricted by orthogonality like PCA, we propose that AA will be able to extract biological information despite the different experimental conditions of the five studies. We show how AA succeeds in identifying genes that undergo changes in gene expression during evolutionary adaptation of the bacteria to the cystic fibrosis lung.

Methods

The diagram in Figure 1 illustrates the process of AA. First data is collected and pre-processed. Pre-processing includes extraction of the raw data *cel*-files in R by use of the package *affy* [13]. Then, data is normalized using the qspline method [14] and gene expression index values are calculated using *robust multiarray average* expression measure [15]. The next stage is to apply the AA algorithm to the expression matrix and calculate explained variance in order to evaluate the solution. Once the archetypes are defined, it is possible to see how samples cluster together based on their relation to the archetypes. Finally, the archetypes can be characterized in a biological context based on their gene expression profiles. The gene expression values were not calculated relative to control strains since different control strains were used across the five analyzed studies.

Data collection

We analyzed cDNA microarray data from four previously published *in vitro* studies (study 1–4) of *P. aeruginosa* sampled from CF lung infections. Three of the data sets were obtained from the online NCBI Gene Expression Omnibus

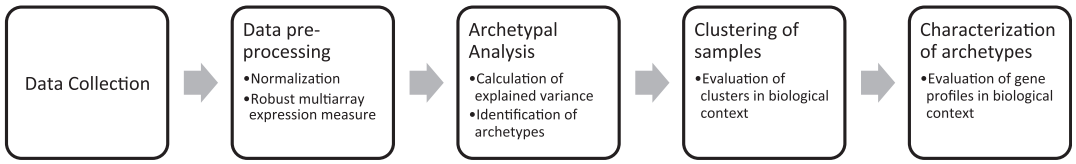


Figure 1 Flow diagram of the archetypal analysis. First, data is collected and pre-processed. Then, Archetypal Analysis is applied resulting in a clustering of samples based on the closest defined archetype. Finally, the archetypes are characterized and evaluated in a biological context.

(GEO) Database with the accession numbers GSE21966 [16], GSE31227 [17] and GSE10362 [18]. The fourth data set by Lee et al. [19] was obtained through request directly to the corresponding author. A fifth data set was generated for this study (study 5). An overview of the microarray data set is shown in Table 1.

Study 1 [16]. This gene expression data set consists of 17 samples (in duplicates) representing clonal isolates sampled from three CF patients on timescales ranging from 3 months to 8 years. Two of the patients each harbored a unique clone (A and B), whereas a strain replacement occurred in the third patient, and two individual clones Ca and Cb were therefore isolated from this patient. For each isolate, information about colony morphology is available, and for the present analysis, we grouped these morphotypes into two categories: Mucoid ('mucoid' morphotypes) and non-mucoid ('dwarf' and 'classic' morphotypes).

The experimental procedures are fully described by Huse et al. [16]. In brief, cells were grown in synthetic cystic fibrosis sputum medium (SCFSM) to an optical density read at 600 nm (OD_{600}) of 0.4-0.5 prior to Affymetrix *P. aeruginosa* GeneChip microarray analysis. The strains *P. aeruginosa* PAO1 and *P. aeruginosa* PA14 (referred to as PAO1 and PA14 respectively) were included as controls in their study. PAO1 was originally isolated from a burn wound [20] and has been widely used as a reference strain for studies of *P. aeruginosa.* PA14 is a highly virulent laboratory strain that most likely represents an environmental strain of *P. aeruginosa*, although it has also been isolated from CF lungs in Europe [21,22].

Study 2 [17]. This data set consists of different clonal lineages isolated from the lungs of CF patients (B6, B12, B38, CF30, CF46, CF66, CF105, CF114, CF173, CF211, CF243, CF333 and CF506) between 1973 and 2008 spanning early stage infection to chronic stage infection [23]. Many of the isolates from study 2 share the same clonal type called "DK2". The data set consists of 29 samples in triplicates. One group of samples was isolated from CF children between 2006 and 2008 and these isolates represent early stage infection. Each isolate was characterized based on two colony morphotypes; mucoid and non-mucoid. The data set includes a sequential mucoid and non-mucoid paired strain, where the non-mucoid strain (B38-2NM) was generated *in vitro* by allelic replacement of its *mucA* allele [24]. Cells were grown in Luria-Bertani (LB) medium to OD_{600} of 0.5 (OD_{600} = 1 for samples #129-140) prior to Affymetrix *P. aeruginosa* GeneChip microarray analysis. PAO1 was included as control in this study.

Study 3 [19]. This data set consists of twelve clonally related, sequential mucoid and non-mucoid paired *P. aeruginosa* isolates. The isolates were obtained from three CF patients. All isolates from study 3 share the same clonal type called "DK1". Cells were grown in beef broth (BB) to an OD_{600} of 1 prior to Affymetrix *P. aeruginosa* GeneChip microarray analysis. Each experiment was done in duplicate. Isolates with high mutation rates (hereafter, "hypermutators") were identified within the data set.

Study 4 [18]. This data set consists of eight sequential isogenic isolates recovered over a period of three to five years from a single CF patient (patient M). The isolates included both hypermutators and non-hypermutators and one isolate was mucoid. Cells were grown in LB medium and harvested during late-logarithmic growth phase at optical density above 3. Each sample was triplicated.

Study 5 (this study). This data set consists of four isolates from the same patient (CF211). The isolates are two mucoid/non-mucoid pairs isolated together in 1997 and 2006 respectively. Cells were grown in BB to an OD_{600} of 1 prior to Affymetrix *P. aeruginosa* GeneChip microarray analysis. Microarray data were generated using Affymetrix protocols as previously described [23]. Each experiment was done in triplicates. The isolates share the same clone type "DK2" as many of the isolates from study 2, but the experimental conditions are similar to those in study 3.

Archetypal analysis

The fundamental principle of AA is briefly introduced below. AA is fully described by Cutler and Breiman [6]. AA is defined by the decomposition

$$\boldsymbol{X} \approx \boldsymbol{XCS},$$
$$s.t. \quad \boldsymbol{C} \geq 0, \sum_{n=1}^{N} \boldsymbol{c}_{nd} = 1, \quad \boldsymbol{S} \geq 0, \sum_{d=1}^{D} \boldsymbol{S}_{dn} = 1.$$

Thøgersen et al. BMC Bioinformatics 2013, 14:279
http://www.biomedcentral.com/1471-2105/14/279

Table 1 List of samples

Sample #	Sample name	Study	Patient	Clone	Year	Mucoid	Mutator	State[1]	Medium[2]	OD[3]
[1,2]	Huse_A1	Study 1	A	"A"	~1983	No	N/A	Early	SCFSM	0.4-0.5
[3,4]	Huse_A2	Study 1	A	"A"	~1984	No	N/A	Early	SCFSM	0.4-0.5
[5,6]	Huse_A3.1 (m)	Study 1	A	"A"	~1985	Yes	N/A	Early	SCFSM	0.4-0.5
[7,8]	Huse_A3.2	Study 1	A	"A"	~1985	No	N/A	Early	SCFSM	0.4-0.5
[9,10]	Huse_A4 (m)	Study 1	A	"A"	~1986	Yes	N/A	Early	SCFSM	0.4-0.5
[11,12]	Huse_B1	Study 1	B	"B"	~1983	No	N/A	Early	SCFSM	0.4-0.5
[13,14]	Huse_B2.1	Study 1	B	"B"	~1987	No	N/A	Late	SCFSM	0.4-0.5
[15,16]	Huse_B2.2	Study 1	B	"B"	~1987	No	N/A	Late	SCFSM	0.4-0.5
[17,18]	Huse_B2.3 (m)	Study 1	B	"B"	~1987	Yes	N/A	Late	SCFSM	0.4-0.5
[19,20]	Huse_B3.1 (m)	Study 1	B	"B"	~1991	Yes	N/A	Late	SCFSM	0.4-0.5
[21,22]	Huse_B3.2	Study 1	B	"B"	~1991	No	N/A	Late	SCFSM	0.4-0.5
[23,24]	Huse_B3.3 (m)	Study 1	B	"B"	~1991	Yes	N/A	Late	SCFSM	0.4-0.5
[25,26]	Huse_Ca1	Study 1	C	"Ca"	~1983	No	N/A	Early	SCFSM	0.4-0.5
[27,28]	Huse_Ca2 (m)	Study 1	C	"Ca"	~1983	Yes	N/A	Early	SCFSM	0.4-0.5
[29,30]	Huse_Cb1 (m)	Study 1	C	"Cb"	~1987	Yes	N/A	Early	SCFSM	0.4-0.5
[31,32]	Huse_Cb2	Study 1	C	"Cb"	~1987	No	N/A	Late	SCFSM	0.4-0.5
[33,34]	Huse_Cb3 (m)	Study 1	C	"Cb"	~1987	Yes	N/A	Late	SCFSM	0.4-0.5
[35-36]	Huse_PA14	Study 1	N/A	"PA14"	N/A	No	N/A	wt	SCFSM	0.4-0.5
[37-38]	Huse_PAO1	Study 1	N/A	"PAO1"	N/A	No	N/A	wt	SCFSM	0.4-0.5
[39-41]	Yang_PAO1	Study 2	N/A	"PAO1"	N/A	No	N/A	wt	LB	0.5
[42-47]	Yang_CF510-2006	Study 2	N/A	"WTB"	N/A	No	N/A	N/A	LB	0.5
[48-50]	Yang_B6.0	Study 2	B6	"B6"	~2005	No	N/A	Early	LB	0.5
[51-53]	Yang_B6.4	Study 2	B6	"B6"	~2007	No	N/A	Early	LB	0.5
[54-56]	Yang_B12.0	Study 2	B12	"B12"	~2005	No	N/A	Early	LB	0.5
[57-59]	Yang_B12.4	Study 2	B12	"B12"	~2007	No	N/A	Early	LB	0.5
[60-62]	Yang_B12.7	Study 2	B12	"B12"	~2009	No	N/A	Early	LB	0.5
[63-65]	Yang_B38.1	Study 2	B38	"B38"	~2005	No	N/A	Early	LB	0.5
[66-68]	Yang_B38.2 (m)	Study 2	B38	"B38"	N/A	Yes	N/A	Early	LB	0.5
[69-71]	Yang_B38.2	Study 2	B38	"B38"	~2005	No	N/A	Early	LB	0.5
[72-74]	Yang_B38.6 (m)	Study 2	B38	"B38"	~2006	Yes	N/A	Early	LB	0.5
[75-77]	Yang_CF43-1973	Study 2	CF43	"DK2"	1973	No	N/A	Early	LB	0.5
[78-80]	Yang_CF66-1973	Study 2	CF66	"DK2"	1973	No	N/A	Late	LB	0.5
[81-83]	Yang_CF105_1973	Study 2	CF105	"DK2"	1973	No	N/A	Early	LB	0.5
[84-86]	Yang_CF114_1973	Study 2	CF114	"DK2"	1973	No	N/A	Early	LB	0.5
[87-89]	Yang_CF30-1979	Study 2	CF30	"DK2"	1979	No	N/A	Late	LB	0.5
[90-92]	Yang_CF173-1984	Study 2	CF173	"DK2"	1984	No	N/A	Late	LB	0.5
[93-95]	Yang_CF333-1991	Study 2	CF333	"DK2"	1991	No	N/A	Late	LB	0.5
[96-98]	Yang_CF66-1992	Study 2	CF66	"DK2"	1992	No	N/A	Late	LB	0.5
[99-101]	Yang_CF333_1997	Study 2	CF333	"DK2"	1997	No	N/A	Late	LB	0.5
[102-104]	Yang_CF173-2002	Study 2	CF173	"DK2"	2002	No	N/A	Late	LB	0.5
[105-107]	Yang_CF243-2002	Study 2	CF243	"DK2"	2002	No	N/A	Late	LB	0.5
[108-110]	Yang_CF333-2003	Study 2	CF333	"DK2"	2003	No	N/A	Late	LB	0.5
[111-113]	Yang_CF173-2005	Study 2	CF173	"DK2"	2005	No	N/A	Late	LB	0.5
[114-116]	Yang_CF333-2005	Study 2	CF333	"DK2"	2005	No	N/A	Late	LB	0.5

Table 1 List of samples *(Continued)*

[117-119]	Yang_CF333-2007.1	Study 2	CF333	"DK2"	2007	No	N/A	Late	LB	0.5
[120-122]	Yang_CF333-2007.2 (m)	Study 2	CF333	"DK2"	2007	Yes	N/A	Late	LB	0.5
[123-125]	Yang_CF333-2007.3 (m)	Study 2	CF333	"DK2"	2007	Yes	N/A	Late	LB	0.5
[126-128]	Yang_CF66-2008	Study 2	CF66	"DK2"	2008	No	N/A	Late	LB	0.5
[129-131]	SD_CF211-1997 (m)	Study 5	CF211	"DK2"	1997	Yes	N/A	Late	LB	1
[132-134]	SD_CF211-1997	Study 5	CF211	"DK2"	1997	No	N/A	Late	LB	1
[135-137]	SD_CF211-2006 (m)	Study 5	CF211	"DK2"	2006	Yes	N/A	Late	LB	1
[138-140]	SD_CF211-2006	Study 5	CF211	"DK2"	2006	No	N/A	Late	LB	1
[141-142]	Lee_CF30-1992 (m)	Study 3	CF30	"DK1"	1973	Yes	No	Late	BB	1
[143-144]	Lee_CF30-1992	Study 3	CF30	"DK1"	1973	No	No	Late	BB	1
[145-146]	Lee_CF30-2001 (m)	Study 3	CF30	"DK1"	2001	Yes	No	Late	BB	1
[147-148]	Lee_CF30-2001	Study 3	CF30	"DK1"	2001	No	No	Late	BB	1
[149-150]	Lee_CF46-1988 (m)	Study 3	CF46	"DK1"	1988	Yes	No	Late	BB	1
[151-152]	Lee_CF46-1988	Study 3	CF46	"DK1"	1988	No	No	Late	BB	1
[153-154]	Lee_CF46-1997 (m) HYP	Study 3	CF46	"DK1"	1997	Yes	Yes	Late	BB	1
[155-156]	Lee_CF46-1997 HYP	Study 3	CF46	"DK1"	1997	No	Yes	Late	BB	1
[157-158]	Lee_CF128-1992 (m)	Study 3	CF128	"DK1"	1992	Yes	No	Late	BB	1
[159-160]	Lee_CF128-1992 HYP	Study 3	CF128	"DK1"	1992	No	Yes	Late	BB	1
[161-162]	Lee_CF128-2002 (m)	Study 3	CF128	"DK1"	2002	Yes	No	Late	BB	1
[163-164]	Lee_CF128-2002 HYP	Study 3	CF128	"DK1"	2002	No	Yes	Late	BB	1
[165-167]	Hob_1998	Study 4	M	"M"	1998	No	No	Late	LB	>3
[168–170]	Hob_1998 (m)	Study 4	M	"M"	1998	Yes	No	Late	LB	>3
[171-173]	Hob_1999	Study 4	M	"M"	1999	No	No	Late	LB	>3
[174–176]	Hob_2001	Study 4	M	"M"	2001	No	No	Late	LB	>3
[177-179]	Hob_1999 HYP	Study 4	M	"M"	1999	No	Yes	Late	LB	>3
[180-182]	Hob_2001.1 HYP	Study 4	M	"M"	2001	No	Yes	Late	LB	>3
[183-185]	Hob_2001.2 HYP	Study 4	M	"M"	2001	No	Yes	Late	LB	>3
[186-188]	Hob_2001.3 HYP	Study 4	M	"M"	2001	No	Yes	Late	LB	>3

[1]Adaptation state (early or late).
[2]Growth medium for the experiments: Synthetic Cystic Fibrosis Sputum Medium (SCFSM), Luria-Bertani Broth (LB), or Beef Broth (BB).
[3]Optical density (OD) at 600 nm during cell harvest.

Where we use the notation $\boldsymbol{S} \geq \boldsymbol{0}$ to denote that the entries of a matrix $\boldsymbol{S}$ are constrained non-negative. The archetypes (components) are given as the columns of the matrix $\boldsymbol{A}$ defined by $\boldsymbol{A} = \boldsymbol{XC}$ such that the columns of **A** are formed by convex combinations of the samples.

A *K*-component AA finds a matrix with elements A_{mk} defining *K M*-dimensional archetypes and each data point can be represented by a convex combination of these archetypes. Each archetype thereby has a specific gene profile that is saved in the *k*'th column of $\boldsymbol{A}$, i.e. $\boldsymbol{a}_k$. The coefficients (α_1, α_2, .., α_K) for a given data point $\boldsymbol{x}_n$ are saved in the nth column of the matrix $\boldsymbol{S}$, i.e. $\boldsymbol{s}_n$, with elements S_{kn}. The values of these coefficients range from 0 to 1 and the sum of the coefficients equals 1.

The AA algorithm as for PCA and k-means attempts to minimize the residual sum of squares (RSS).

$$RSS = \sum_{m=1,n=1}^{M,N} (\boldsymbol{X}\text{-}\boldsymbol{AS})_{mn}^2 = \|\boldsymbol{X}\text{-}\boldsymbol{AS}\|_F^2$$

Where M is the number of attributes and N the number of observations.

Determining the characteristics of each of the archetypes can clarify the features of the data set.

Principal component analysis and k-means clustering

Principal component analysis and k-means clustering were applied to the same data set.

Principal component analysis is given by the decomposition

$$\begin{aligned} &\boldsymbol{X} \approx \boldsymbol{AS}, \\ s.t.\ &\boldsymbol{A}^T\boldsymbol{A}=I, \quad \boldsymbol{SS}^T=\boldsymbol{D}, \end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{D}$ is a diagonal matrix with the elements in the diagonal are sorted according to their magnitude.

In k-means clustering $\boldsymbol{S}$ is constrained to be a binary assignment matrix such that $\boldsymbol{A} = \boldsymbol{XS}^{\boldsymbol{T}}\ (\boldsymbol{SS}^{\boldsymbol{T}})^{-1}$ represents the Euclidean centers of each cluster.

Number of components

For AA, it is necessary to set the number of components prior to analysis similar to k-means clustering. Our choice of archetype component was guided by plotting the explained variance as a function of number of components (Figure 2). For the purpose of this study, we chose to analyze seven components, which kept the number of components at a minimum while at the same time explaining a large part (59.3%) of the variance. The standard deviation between 10 repeated iterations is very low, which suggests that the solution is robust. The explained variance for PCA and k-means clustering with seven components were 54.4% and 68.4% respectively. As expected, the PCA model, which is the most flexible of the considered models, has a higher explained variance than AA that in turn has a higher explained variance than the more restricted k-means clustering.

As a quality measure the deviation between the n^{th} original data point $\boldsymbol{x}_n$ and the derived data point $\boldsymbol{XCs}_n$ based on the seven archetypes was calculated. The measure is given as the Explained Sample Variance $\left(ESV = \cdot \frac{\|\boldsymbol{x}_n\|_F^2 - \|\boldsymbol{x}_n - \boldsymbol{XCs}_n\|_F^2}{\|\boldsymbol{x}_n\|_F^2} \cdot\right)$ ranging between 0 and 1 where 1 is a perfect match. By evaluating these *ESV* values, it is possible to state which data points are well described by the model. No conclusions should be made for data points where *ESV* is low, because these data points are poorly described by the model.

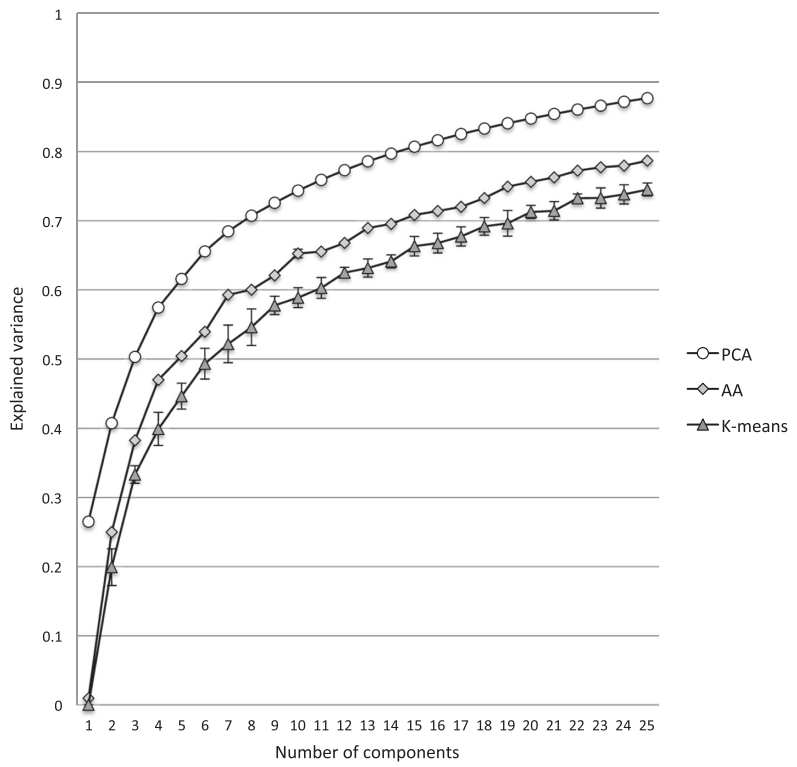


Figure 2 Explained variance. The explained variance plotted as a function of number of components for principal component analysis (PCA) archetypal analysis (AA) and k-means clustering (K-means). The plotted values are the mean of 10 repeated iterations. The standard deviations are indicated with error bars for k-means clustering. The standard deviations for archetypal analysis are very small and therefore not visible.

Characterization of archetypes

Each archetype was characterized based on its specific gene profile. This was done by identifying genes with statistically significant transcriptional changes. Genes with more than a two-fold change in expression value, compared to the mean expression of the respective gene for all samples, were indicated as up-regulated whereas genes with less than 0.5-fold change were indicated as down-regulated. Genes were assigned to 26 different gene ontology (GO) classes based on the gene annotation for *P. aeruginosa* PAO1 from the Pseudomonas Genome Database [25]. If a gene was assigned to more than one GO class it was re-assigned to the most overall GO class (Additional file 1: Table S1). GO classes that were over-represented within the group of up-or down-regulated genes were identified by the Hypergeometric distribution test with significance level p = 0.01 [26].

Matlab code

The methods mentioned above were implemented in Matlab unless otherwise stated. The Matlab Code for AA is available online at http://www.mortenmorup.dk. This code estimates $\boldsymbol{C}$ and $\boldsymbol{S}$ using a projected gradient descent iterative approach initialized by the FurthestSum procedure, for details see also [5]. A brief description of the Matlab Code is listed in Additional file 2.

Results and discussion

To explore the value of Archetypal Analysis in gene expression studies, we assembled microarray data from five separate studies of clinical *P. aeruginosa* sampled from CF lung infections [16-19]. These studies measured global gene expression of different clonal *P. aeruginosa* isolates under diverse *in vitro* growth conditions. The studied bacterial isolates exhibited different clinically relevant phenotypes such as mucoidy and hypermutability, were different clone types, and were isolated from patients at different time points in relation to disease progression (Table 1).

Defining archetypes in the data set

AA was performed on a data set with 188 samples in total (sum of duplicates and triplicates) using the code provided in [5] and additional codes that are available online (see Methods section). Seven archetypes were identified for the integrated data set. The contribution of the individual archetypes to each sample is visualized as a heat map of the coefficient matrix $\boldsymbol{S}$ in Figure 3. The relation between the

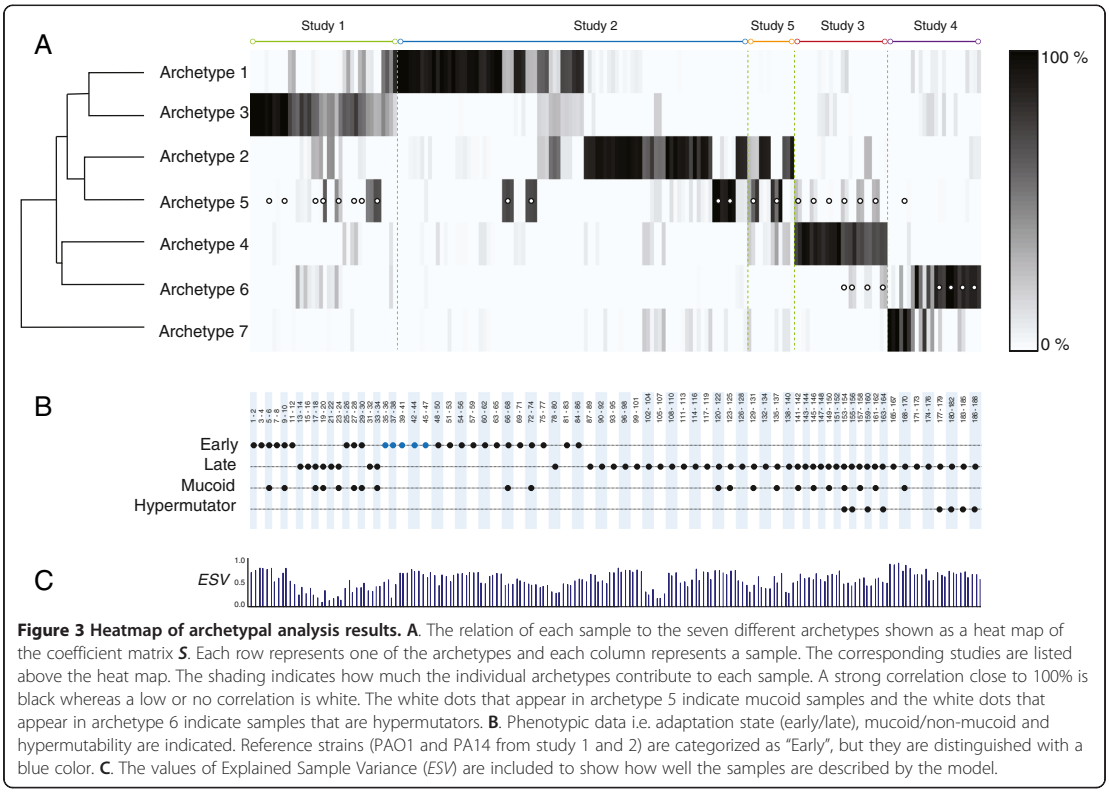


Figure 3 Heatmap of archetypal analysis results. A. The relation of each sample to the seven different archetypes shown as a heat map of the coefficient matrix $\boldsymbol{S}$. Each row represents one of the archetypes and each column represents a sample. The corresponding studies are listed above the heat map. The shading indicates how much the individual archetypes contribute to each sample. A strong correlation close to 100% is black whereas a low or no correlation is white. The white dots that appear in archetype 5 indicate mucoid samples and the white dots that appear in archetype 6 indicate samples that are hypermutators. **B**. Phenotypic data i.e. adaptation state (early/late), mucoid/non-mucoid and hypermutability are indicated. Reference strains (PAO1 and PA14 from study 1 and 2) are categorized as "Early", but they are distinguished with a blue color. **C**. The values of Explained Sample Variance (*ESV*) are included to show how well the samples are described by the model.

gene profiles of the seven archetypes is shown in a dendrogram based on hierarchical clustering (Figure 3).

Archetypal analysis separates study 2 into two groups representing adapted and non-adapted isolates respectively

Archetype 1, 2 and 5 represent samples from study 2. This appears from Figure 3 by these samples having coefficients close to one (100%) in one of the three archetypes. Study 2 is composed of samples that were isolated from cystic fibrosis patients from the Danish CF clinic between 1973 and 2008 [17]. The samples from this study can be divided into two groups; one group representing isolates from early infection (hereafter referred to as 'non-adapted' isolates), and one group representing isolates from long-term chronic infection (hereafter referred to as 'adapted' isolates). Archetype 1 represents non-adapted isolates from study 2 including the reference strain PAO1 and an isolate called CF510-2006 that is considered as an ancestor to many of the isolates from study 2 [27]. CF510-2006 has phenotypic characteristics similar to wild type environmental *P. aeruginosa* strains [27]. Archetype 2 represents adapted isolates from study 2. One of the isolates from study 2 (triplicate samples #78-80) is best explained by archetype 2, although it was isolated in 1973 and is considered as an early isolate with respect to time of isolation. However, from genomic studies of study 2 it is known that this isolate has two mutations located in the genes *rpoN* and *mucA* and these mutations are common to the adapted isolates and they are associated with an adapted phenotype [23]. This isolate therefore can justifiably be considered to belong to the group of adapted isolates. The archetypes 1 and 2 thereby successfully cluster study 2 into two distinct groups based on adaptation level.

Some of the samples from both groups of study 2 are also, to a greater or lesser degree, based on archetype 5. These samples all have a mucoid phenotype characterized by an over-production of alginate. The transition from a non-mucoid to a mucoid phenotype is often observed during adaptation of the bacteria to the CF lung and this shift is important for establishment of chronic infections [28].

Characterization of archetype 1, 2 and 5

We next studied the up-and down-regulated genes within each archetype to find patterns that would suggest specific biological properties associated with archetype 1, 2 and 5. Figure 4 shows the distribution of significantly up- and down-regulated genes with respect to GO classes for these three archetypes. GO classes that were over-represented within the group of up-or down-regulated genes were identified by Hypergeometric distribution test [26].

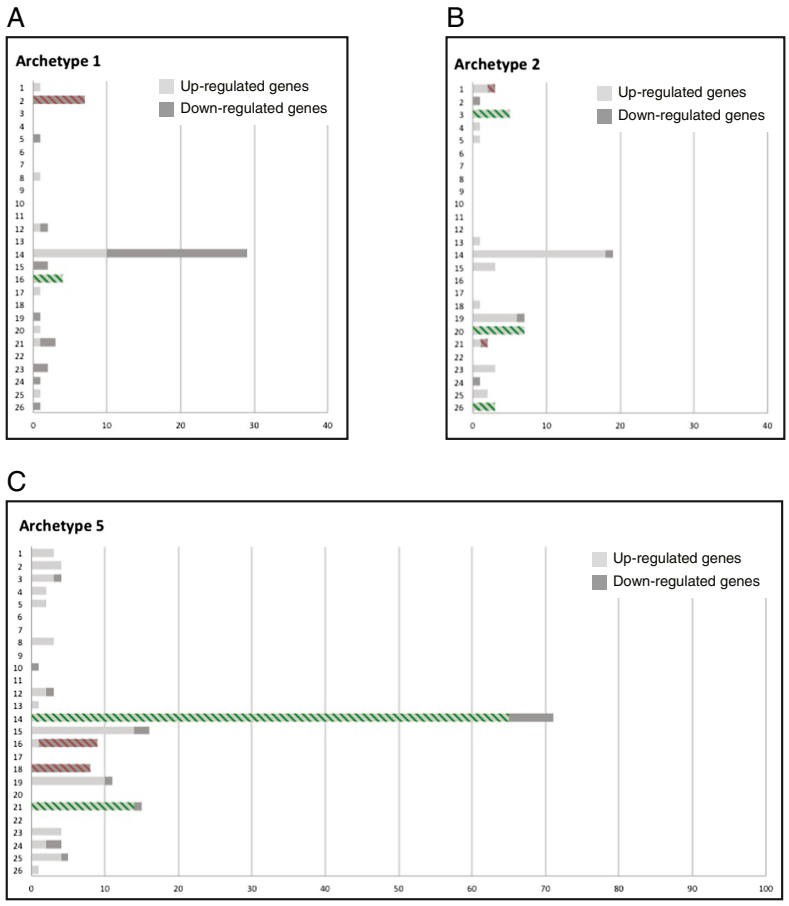
A full list of up-and down-regulated genes for all archetypes can be found in Additional file 1: Table S1. From the archetype characterization in Figure 4A, it appears that the early strains represented by archetype 1 have a high expression of genes belonging to the GO class "Motility and Attachment". At the same time, they have a low expression of genes related to "Amino acid biosynthesis and metabolism". The adapted strains represented by archetype 2 are characterized by up-regulation of genes related to "Antibiotic resistance and susceptibility", "Two-component regulatory systems" and genes "Related to phage, transposon and plasmid" (Figure 4B). Down-regulated genes belong to the functional classes "Adaptation, Protection" and "Secreted factors". These observations are in agreement with earlier studies examining the phenotypic differences between non-adapted and adapted isolates [17,23,24,29]. Archetype 5 was primarily characterized by a strong up-regulation of genes related to alginate biosynthesis belonging to the GO class "Secreted factors" (Figure 4C). This is in agreement with the mucoid phenotype, characterized by overproduction of alginate that is observed for all the samples that have an apparent coefficient for this archetype. This archetype is also characterized by up-regulation of many genes encoding hypothetical proteins and down-regulation of genes involved in "Motility and Attachment" and "Protein secretion".

AA succeeds in clustering study 2 into biologically meaningful groups. At the same time, it is easy to extract biological features important for all groups. So far, the AA analysis is verified since the characteristics of the archetypes 1, 2 and 5 are consistent with results from genotypic and phenotypic studies of study 2 [23].

The identification of these genes thereby validates this model and we are able to find biological characteristics of the different samples by analyzing the archetypes. For each of the archetypes the lists of up-and down-regulated genes also include genes encoding hypothetical proteins and it is possible that such genes are also involved in the adaptation process. For archetype 5 there are a large proportion of up-regulated genes belonging to the GO class "hypothetical proteins". Further experimental studies are required to understand the function of these genes and their relation to the adaptation process.

Parallel adaptation processes are observed between study 1 and study 2

Archetype 3 is defined close to a subset of samples (#1-10) from study 1. These samples all have the same genotype (clone A) and they are considered as non-adapted since they were isolated early during the infection history of the patient (cf. Table 1) [16]. This archetype is characterized by up-regulation of genes belonging to the GO classes "Motility and attachment", "Protein secretion" and "Secreted factors" and many of these genes are related to type III secretion and pilin biosynthesis. Archetype 3 is characterized by down-regulation of "Antibiotic resistance



1 Adaptation, Protection
2 Amino acid biosynthesis and metabolism
3 Antibiotic resistance and susceptibility
4 Biosynthesis of cofactors, prosthetic groups and carriers
5 Carbon compound metabolism
6 Cell division
7 Cell wall / LPS / capsule
8 Central intermediary metabolism
9 Chaperones & heat shock proteins
10 Chemotaxis
11 DNA replication, recombination, modification and repair
12 Energy metabolism
13 Fatty acid and phospholipid metabolism
14 Hypothetical, unclassified, unknown
15 Membrane proteins
16 Motility and attachment
17 Nucleotide biosynthesis and metabolism
18 Protein secretion / export apparatus
19 Putative enzymes
20 Related to phage, transposon or plasmid
21 Secreted factors (toxins, enzymes, alginate)
22 Transcription, RNA processing and degradation
23 Transcriptional regulation
24 Translation, post-translational modification, degradation
25 Transport of small molecules
26 Two-component regulatory systems

Figure 4 Characterization of archetype 1, 2 and 5. Number of up-and down-regulated genes within 26 gene ontology classes for archetype 1 **(A)**, archetype 2 **(B)** and archetype 5 **(C)**. Enriched gene-ontology classes are cross-hatched in green and red for up-and down-regulated genes respectively. The values on the x-axes are number of genes.

and susceptibility" and genes "Related to phage, transposon and plasmid" (Figure 5A).

Archetype 1 and 3 represent early isolates from study 2 and study 1 respectively. The two archetypes share characteristics with respect to up-regulation of "Motility and attachment" and down-regulation of genes with relation to "Adaptation and antibiotic resistance". Hierarchical clustering of the seven archetypes also groups archetype 1 and 3 together shown in a dendrogram in Figure 3A.

Samples #11-15 are the earliest isolates of another clone (clone B) from study 1 and they are also closely related to archetype 3. Samples #25-30 represent early

Thøgersen *et al. BMC Bioinformatics* 2013, **14**:279
http://www.biomedcentral.com/1471-2105/14/279

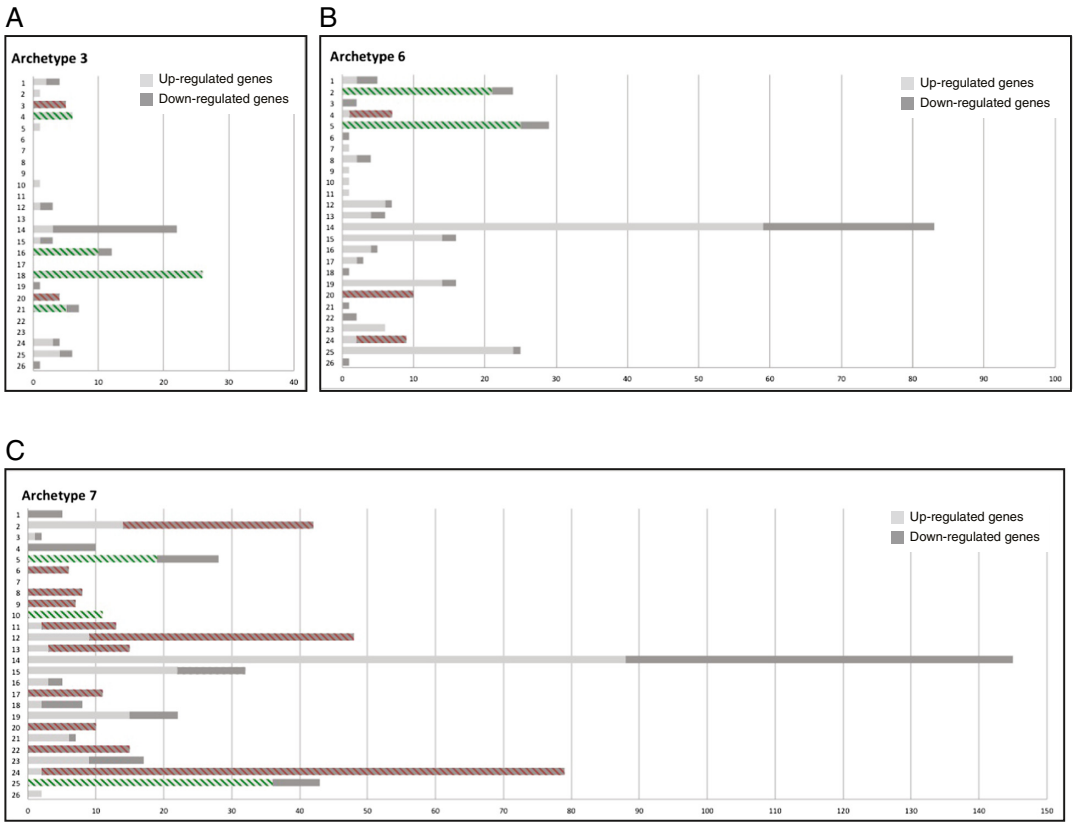


Figure 5 Characterization of archetype 3, 6 and 7. Number of up-and down-regulated genes within 26 gene ontology classes for each archetype for archetype 3 **(A)**, archetype 6 **(B)** and archetype 7 **(C)**. Enriched gene-ontology classes are cross-hatched in green and red for up-and down-regulated genes respectively. The values on the x-axes are number of genes.

isolates of two other clones (clone Ca and Cb) from study 1 and they are best described by archetype 3. However, they also show recognizable coefficients (weak bands in Figure 3A) for archetype 1, which also indicates the similarity between the non-adapted samples from study 1 and study 2. This also applies to samples #9-10 that are the latest isolate of the clonal group A from study 1. The reference strains PA14 and PAO1 are included in study 1 and they are best described by archetype 1 that also represents PAO1 samples from study 2. Differences between data from study 1 and study 2 are therefore most likely to be due to different clonal lineages more than experimental differences. Samples #17-24 are late isolates of clone B from study 1. Unfortunately, the samples are poorly described by the model, as indicated by the low *ESV* values. However, the samples show similarity to archetype 2 representing the adapted isolates from study 2. Together these findings suggest that the adaptation processes from the two studies 1 and 2 are parallel.

Samples #5-6, #9-10, and #27-30 are reported as mucoid but they do not appear to be similar to the mucoid isolates from study 2, where archetype 5 identified all the

mucoid isolates. However, archetype 5 succeeds in identifying the mucoid isolates for samples #20, #23-24 (weak indication) and #33-34.

Archetypal analysis groups the samples from study 5 together with its clonal relatives from study 2 despite different experimental conditions

Archetype 2 and 5 best describe study 5. The samples that have a coefficient close to one for archetype 5 are mucoid and this is consistent to the results for study 2. The non-mucoid isolates are close to archetype 2, which represents non-mucoid isolates from study 2 sharing the same clonal type as the isolates in study 5. This shows a strong consistency between study 2 and 5, although study 5 was performed under experimental conditions similar to those in study 3.

The five analyzed studies were performed under diverse experimental conditions including different media types. We compare the characteristics of the seven described archetypes and some of the differences are most likely due to the effect of the different media. This study does not account for how the different media alone affect the transcriptome. However, when we compare the different archetypes we have seen that the samples cluster more into groups of clonally related bacteria than into clusters of samples exposed to the same experimental procedure e.g. PAO1 in study 1 and study 2 and the samples from study 2 and study 5. The effect of the diverse media types does therefore not override the real biological relation between the bacteria and we justify comparing the samples from the five studies despite different experimental procedures. Future investigations of clonally related bacteria may further examine the effect from the media alone on the transcriptome.

A single archetype represents hypermutators for study 3 and study 4

Archetype 4 mainly represents study 3. Study 3 is also composed of samples derived from the Danish CF clinic representing adapted isolates as for study 2. However, the samples share another clonal type (DK1) and the experiments are performed under different conditions than those used for study 2. The differences between archetype 2 and 4 are most likely due to clonal differences more than experimental differences since the same differences in experimental conditions did not separate study 2 and study 5 in this analysis. A plot of enriched gene ontology classes for archetype 4 similar to plots in Figure 4 is accessible in Additional file 3. The samples from study 3 differ from each other as some of them have a minor recognizable coefficient in archetype 5 or archetype 6. Archetype 5 represented the mucoid isolates from study 2. All the samples from study 3 that have a recognizable coefficient for archetype 5 are in fact mucoid. In this case, knowledge from one study can be transferred to another study despite the different experimental conditions and clonal types between the two studies. Archetype 6 represents samples from study 4. The samples that are closest to this archetype are all hypermutators.

The samples from study 3 with recognizable coefficients for archetype 6 are also hypermutators. One of the hypermutator samples in study 3 is not identified as having a recognizable coefficient for archetype 6. However, this sample stands out from the rest of the hypermutators by also being mucoid. The analysis thereby suggests an archetype that is able to characterize hypermutators in general. The similarity of the hypermutators could be due to similar selective pressures present in the lung environments of CF patients. This analysis could suggest that the hypermutators follow the same path of evolution despite many changes arising as a consequence of mutations.

Hypermutation is often due to mutations in the *mutS* or *mutL* genes that are part of the mismatch repair system [30]. The hypermutator trait is often observed for adapted strains of *P. aeruginosa* [18,19,31,32] and the high mutation rate is thought to be advantageous in the changing host environment due to acceleration of adaptation [18,30]. A reason for the hypermutators to develop a similar adaptive phenotype, but different from the adapted non-mutators, could be the chance of obtaining a combination of multiple adaptive mutations at one time, which is less likely for strains with a normal mutation rate [32]. Another possibility is that the *mutS* gene or the *mutL* gene possesses a regulatory function that is altered due to the mutation in the respective gene. There is some evidence that bacteria can sense the missing mismatch repair function and this will influence transcriptional regulation [33]. This would make a fingerprint on the gene expression profiles for the hypermutators resulting in similarity between the gene expression profiles. A third possibility is that the mutation targets of the hypermutators are biased due to for example a preference of specific transversions and transitions and other phenomena [34]. This analysis suggests that there is a common phenotypic trait between the hypermutators. The underlying reason needs further investigation.

Amino acid biosynthesis and metabolism are important for adaptation to the cystic fibrosis lung

The characteristics of archetype 6 might be used to better understand the features shared by the hypermutators. However, the experimental procedure used for study 4 is markedly changed since the samples are harvested in late-logarithmic growth phase (optical density read at 600 nm ≥ 3) compared with exponential growth conditions for study 1, 2, 3 and 5. The observed up-and down-regulated genes can therefore be ascribed to changes

Thøgersen *et al. BMC Bioinformatics* 2013, **14**:279
http://www.biomedcentral.com/1471-2105/14/279

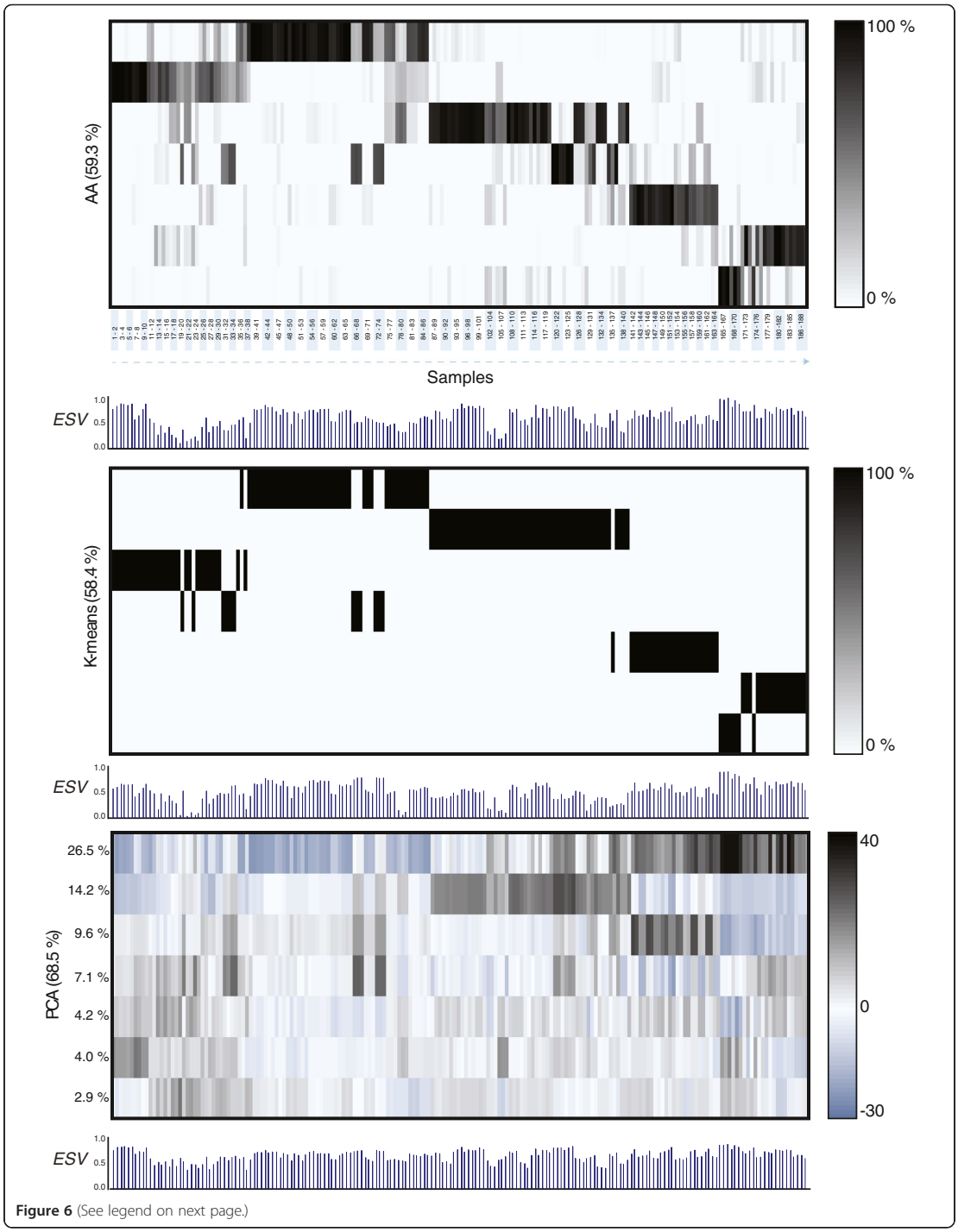


Figure 6 (See legend on next page.)

(See figure on previous page.)
Figure 6 Comparison between archetypal analysis, principal component analysis and k-means clustering. Visual representation of a seven-component analysis using archetypal analysis (AA), principal component analysis (PCA) and k-means clustering (K-means). Explained sample variance (*ESV*) for each analysis is included. For each PCA component the contribution to explained variance is indicated. The explained variance for a seven component analysis is indicated in brackets for each analysis.

during the transition from exponential to stationary growth phase more than to changes due to accumulated mutations. In order to exclude effects due to the growth phase, archetype 6 is compared to archetype 7. Both archetypes represent samples from study 4. Archetype 7 is mainly represented by non-hypermutators that constitute isogenic pairs to the samples represented by archetype 6.

For archetype 7 many GO classes are overrepresented by either up-or downregulated genes (Figure 5C). This is most likely due to the different growth conditions in study 4 compared to the other four studies. The profile of archetype 7 is very different from the other studies suggesting significant changes in the transcriptome due to the change in growth conditions. If we consider archetype 6 (Figure 5B), we do not observe the same dramatic changes. This can also be seen from the dendrogram in Figure 3 where archetype 6 is closer to the remaining archetypes than archetype 7. This could indicate that the hypermutators represented by archetype 6 are not that sensitive to the changes in growth conditions compared to the non-mutators, represented by archetype 7. An explanation for this could be that the hypermutators possess mutations in regulatory genes that make the gene expression less sensitive to the surrounding conditions, in this case growth phase and cell density.

Archetype 6 is also characterized by up-regulation of genes involved in the GO-class 2 ('Amino acid biosynthesis and metabolism') and down-regulation of genes from GO-class 4 ('Biosynthesis of cofactors, prosthetic groups and carriers') suggesting that these are important during adaptation of *P. aeruginosa* to the CF lung for the hypermutators.

These findings are to a certain extent similar to what Hoboth et al. [18] found when comparing hypermutator isolates with a non-mutator isolate. They also found amino acid biosynthesis and metabolism to be important for adaptation together with other metabolic pathways [18]. One difference in the comparison is that they compared

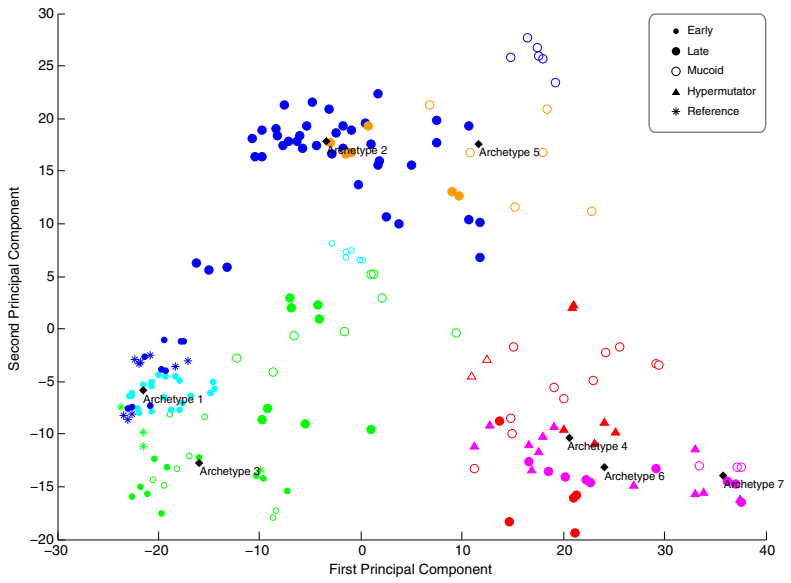


Figure 7 Principal component analysis scatter plot. Each sample is plotted with respect to the loadings of first and second PCA component. The seven archetypes from archetypal analysis are transformed into the PCA space through a basis transformation. Each Study is indicated with a specific color. Study 1: GREEN, study 2: CYAN (samples #48-74) and BLUE (samples #75-128), study 3: RED, study 4: MAGENTA, study 5: ORANGE. The phenotypes are indicated with symbols as "Early", "Late", "Mucoid" and "Hypermutator". The reference strains PAO1 and PA14 from study 1 and 2 are indicated with a symbol as "Reference".

the transcriptomic profiles directly, whereas in this analysis the two proposed archetypal gene expression profiles for archetype 6 and 7 are compared, where archetype 6 represents hypermutators and archetype 7 represents non-mutators. We suggest that the characteristics of the archetype 6 are representative for general hypermutator characteristics, since archetype 6 accounts for hypermutators across different clonal types and across different experimental conditions (study 3 and 4). The gene expression profile of archetype 6 therefore most likely can be linked to the hypermutator trait and its influence on adaptation in the CF lung.

Archetypal analysis supplements principal component analysis and k-means clustering

Results of k-means clustering and PCA of the data set are illustrated together with the results of AA in Figure 6. The results of k-means clustering show how samples are divided into seven groups. The clustering pattern is similar to the pattern from AA but each sample is assigned to only one cluster making k-means clustering rigid compared to AA.

PCA captures most of the explained variance in the first three components (50.3%). However, the components do not give an apparent grouping of samples in Figure 6. PCA solutions are often visualized by plotting the first two components in a two-dimensional scatter plot as shown in Figure 7. Together the first two components account for 40.7% of the variance present in the data set. From the scatter plot it is neither possible to see any grouping correlated to the mucoid phenotype nor the hypermutator phenotype as identified by AA. This illustrates the value of AA compared to PCA. For the present analysis we were fortunate to know some phenotypic traits (mucoidy and hypermutability) of the samples in the data set. These properties were captured by AA. Even if this information was not available it would still be possible to suggest similarities within the data set based on AA. A drawback of AA and k-means compared to PCA is that the choice of the number of components influences how the components are defined while the iterative estimation procedures for extracting the components may terminate at suboptimal solutions. As the archetypes are constrained to be convex combinations of the observations AA relies on the presence of observations that well represent the distinct aspects in the data.

Conclusions

This is the first time Archetypal Analysis has been applied to analysis of gene expression data. Seven archetypes were able to extract the main characteristics of the dataset. The results show that Archetypal Analysis is successful in clustering of data into biologically meaningful groups. At the same time, the analysis is strengthened by matrix factorization making it possible to describe data points as a combination of archetypes.

Archetype 1 and 2 represent non-adapted and adapted isolates respectively, and characterization of the two archetypes identifies the main changes during adaptation of the bacteria to the CF lung. In this study, it is shown that one archetype represents a group of hypermutators (result of clustering) and other data points share characteristics with this group (result of factorization) enabling identification of hypermutators from another group. The analysis provides results that are easy to interpret and we suggest that this analysis could be used to supplement current methods of gene expression analysis.

Availability of supporting data

The Matlab code for our method is freely available online on the website http://www.mortenmorup.dk.

Additional files

Additional file 1: Table S1. List of up-and down-regulated genes for each archetype.

Additional file 2: Short description of Matlab scripts.

Additional file 3: Enriched gene ontology classes for archetype 4.

Competing interests

The authors declared that they have no competing interests.

Authors' contributions

JCT participated in the design of the study, performed the statistical analysis and wrote the manuscript. MM participated in the statistical analysis and helped to draft the manuscript. SD designed and performed DNA microarray experiments. LJ and SM participated in the design of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the Danish Research Agency and the Lundbeck Foundation to S.M. The Villum Foundation supports work in the laboratory of L.J.

The authors thank Henrik Bjørn Nielsen (Center for Biological Sequence Analysis, Technical University of Denmark) for helpful discussions. We also thank Oana Ciofu (Department of International Health, Immunology and Microbiology, Copenhagen University) for providing the raw gene expression data files from study 3 [19].

Author details

[1]Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark. [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Lyngby, Denmark.

Received: 18 December 2012 Accepted: 3 September 2013
Published: 23 September 2013

References

1. Liu W, Wang B, Glassey J, Martin E, Zhao J: **A novel methodology for finding the regulation on gene expression data.** *Proc Natl Acad Sci U S A* 2009, **19:**267–272.
2. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data.** *Proc Natl Acad Sci* 2001, **98:**10781.

3. Kim MH, Seo HJ, Joung JG, Kim JH: **Comprehensive evaluation of matrix factorization methods for the analysis of DNA microarray gene expression data.** *BMC bioinformatics* 2011, **12**(Suppl 1):S8.
4. Quackenbush J: **Computational analysis of microarray data: nature reviews.** *Genetics* 2001, **2:**418–427.
5. Mørup M, Hansen LK: **Archetypal analysis for machine learning and data mining.** *Neurocomputing* 2012, **80:**54–63.
6. Cutler A, Breiman L: **Archetypal analysis.** *Technometrics* 1994, **36:**338–347.
7. Marinetti S, Finesso L, Marsilio E: **Archetypes and principal components of an IR image sequence.** *Infrared Phys Technol* 2007, **49:**272–276.
8. Porzio GC, Ragozini G, Vistocco D: **On the use of archetypes as benchmarks.** *Appl Stochastic Models Bus Indu* 2008, **24:**419–437.
9. Huggins P, Pachter L, Sturmfels B: **Toward the human genotope.** *Bull Math Biol* 2007, **69:**2723–2735.
10. Schwartz R, Shackney SE: **Applying unmixing to gene expression data for tumor phylogeny inference.** *BMC Bioinforma* 2010, **11:**42.
11. Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U: **Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space.** *Science* 2012, **1157:**1157–1160.
12. Noor E, Milo R: **Efficiency in evolutionary trade-offs.** *Science* 2012, **336:**1114–1115.
13. Gautier L, Cope L, Bolstad BM, Irizarry RA: **Affy–analysis of affymetrix genechip data at the probe level.** *Bioinformatics (Oxford, England)* 2004, **20:**307–315.
14. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild H, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome Biol* 2002, **3:**1–16.
15. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics (Oxford, England)* 2003, **4:**249–264.
16. Huse H, Kwon T, Zlosnik J, Speert D: **Parallel evolution in pseudomonas aeruginosa over 39,000 generations in vivo.** *MBio* 2010, **1:**0–8.
17. Yang L, Rau MH, Yang L, Hoiby N, Molin S, Jelsbak L: **Bacterial adaptation during chronic infection revealed by independent component analysis of transcriptomic data.** *BMC Microbiol* 2011, **11:**184.
18. Hoboth C, Hoffmann R, Eichner A, Henke C, Schmoldt S, Imhof A, Heesemann J, Hogardt M: **Dynamics of adaptive microevolution of hypermutable pseudomonas aeruginosa during chronic pulmonary infection in patients with cystic fibrosis.** *J Infect Dis* 2009, **200:**118–130.
19. Lee B, Schjerling CK, Kirkby N, Hoffmann N, Borup R, Molin S, Høiby N, Ciofu O: **Mucoid pseudomonas aeruginosa isolates maintain the biofilm formation capacity and the gene expression profiles during the chronic lung infection of CF patients.** *APMIS* 2011, **119:**263–274.
20. Holloway BW, Krishnapillai V, Morgan AF: **Chromosomal genetics of pseudomonas.** *Microbiol Rev* 1979, **43:**73–102.
21. Rahme LG, Stevens EJ, Wolfort SF, Shao J, Tompkins RG, Ausubel FM: **Common virulence factors for bacterial pathogenicity in plants and animals.** *Science (New York, N.Y.)* 1995, **268:**1899–1902.
22. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Köhler T, Van Delden C, Weinel C, Slickers P, Tümmler B: **Population structure of pseudomonas aeruginosa.** *Proc Natl Acad Sci U S A* 2007, **104:**8101–8106.
23. Yang L, Jelsbak L, Marvig RL, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK, Ciofu O, Høiby N, Sommer MOA, Molin S: **Evolutionary dynamics of bacteria in a human host environment.** *Proc Natl Acad Sci U S A* 2011, **108:**7481–7486.
24. Rau MH, Hansen SK, Johansen HK, Thomsen LE, Workman CT, Nielsen KF, Jelsbak L, Høiby N, Yang L, Molin S: **Early adaptive developments of pseudomonas aeruginosa after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts.** *Environ Microbiol* 2010, **12:**1643–1658.
25. Winsor GL, Lam DKW, Fleming L, Lo R, Whiteside MD, Yu NY, Hancock REW, Brinkman FSL: **Pseudomonas genome database: improved comparative analysis and population genomics capability for pseudomonas genomes.** *Nucleic Acids Res* 2011, **39:**D596–D600.
26. Rivals I, Personnaz L, Taing L, Potier M: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2007, **23:**401–407.
27. Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L: **Deletion and acquisition of genomic content during early stage adaptation of pseudomonas aeruginosa to a human host environment.** *Environ Microbiol* 2012, **14:**2200–2211.
28. Govan JR, Deretic V: **Microbial pathogenesis in cystic fibrosis: mucoid pseudomonas aeruginosa and burkholderia cepacia.** *Microbiological reviews* 1996, **60:**539–574.
29. Jelsbak L, Johansen HK, Frost AL, Thøgersen R, Thomsen LE, Ciofu O, Yang L, Haagensen JAJ, Høiby N, Molin S: **Molecular epidemiology and dynamics of pseudomonas aeruginosa populations in lungs of cystic fibrosis patients.** *Infect Immun* 2007, **75:**2214–2224.
30. Sundin GW, Weigand MR: **The microbiology of mutability.** *FEMS Microbiol Lett* 2007, **277:**11–20.
31. Mena A, Smith EE, Burns JL, Speert DP, Moskowitz SM, Perez JL, Oliver A: **Genetic adaptation of pseudomonas aeruginosa to the airways of cystic fibrosis patients is catalyzed by hypermutation.** *J Bacteriol* 2008, **190:**7910–7917.
32. Oliver A, Mena A: **Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance.** *Clin Microbiol Infect* 2010, **16:**798–808.
33. Fukui K, Wakamatsu T, Agari Y, Masui R, Kuramitsu S: **Inactivation of the DNA repair genes mutS, mutL or the anti-recombination gene mutS2 leads to activation of vitamin B1 biosynthesis genes.** *PloS one* 2011, **6:**e19053.
34. Moyano AJ, Smania AM: **Simple sequence repeats and mucoid conversion: biased mucA mutagenesis in mismatch repair-deficient pseudomonas aeruginosa.** *PloS one* 2009, **4:**e8203.

doi:10.1186/1471-2105-14-279
Cite this article as: Thøgersen *et al.*: Archetypal analysis of diverse *Pseudomonas aeruginosa* transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics* 2013 **14**:279.

Paper 1: Additional files 1-3

Paper 1 - Additional file 1: List of up-and down-regulated genes for each archetype

(Table S1)

Available online at: http://www.biomedcentral.com/1471-2105/14/279/additional

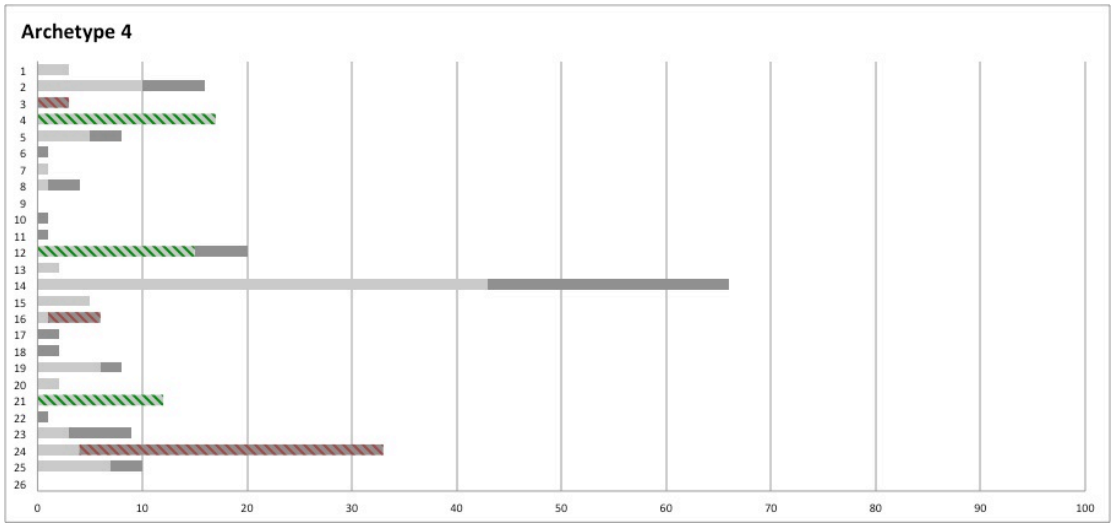
Paper 1 - Additional file 2: Short description of Matlab scripts

GeneArc.m
The script *GeneArc.m* imports microarray data from a specified text file. This text file contains the expression matrix with the gene expression values for all samples. The number of components for the analysis should be given as input before. Archetypal analysis of the data set is executed. The *A* and *S* matrices are computed using the function PCHA and the explained variance for the analysis is calculated. PCHA was described by Mørup and Hansen (2011), and Matlab code is available. The matrix *S* is illustrated as a heat map, which allows easy detection of sample clusters. Results from Principal component analysis (PCA) and K-means clustering are also displayed as heat maps for comparison.

Varexp.m
This script is useful for determination of number of components. *Varexp.m* calculates the explained variance for a 1 to *k* component archetypal analysis and displays explained variance as a function of number of components. The maximal number of components, *k,* must be given as input. The results are displayed as an average of a specified number of runs. Similar plots for PCA and k-means clustering are made. For k-means clustering a high value of *k* can result in empty clusters, which will interrupt the calculations. In this case, a different value of *k* can be defined for k-means clustering and the calculations can proceed.

GeneList.m
The script *GeneList.m* finds significantly up- and down-regulated genes for each archetype compared to the mean values of all samples. The archetypes must be defined using the script *GeneArc.m* before this script is applied. Genes that are found significantly up-regulated are saved in a matrix "Genes_high" whereas genes that are significantly down-regulated are saved in a matrix "Genes_low". In addition to this, information for each gene can be extracted from a specified data file including gene annotation, pathways and functional classes belonging to each gene. This information is saved in two matrices called "Genes_high_list" and "Genes_low_list" for up- and down-regulated genes respectively. The log2 expression values of these genes are listed in two matrices called "Genes_value_high" and "Genes_value_low".

Paper 1 - Additional file 3: Enriched gene ontology classes for archetype 4



1 Adaptation, Protection
2 Amino acid biosynthesis and metabolism
3 Antibiotic resistance and susceptibility
4 Biosynthesis of cofactors, prosthetic groups and carriers
5 Carbon compound catabolism
6 Cell division
7 Cell wall / LPS / capsule
8 Central intermediary metabolism
9 Chaperones & heatshock proteins
10 Chemotaxis
11 DNA replication, recombination, modification and repair
12 Energy metabolism
13 Fatty acid and phospholipid metabolism
14 Hypothetical, unclassified, unknown
15 Membrane proteins
16 Motility & Attachment
17 Nucleotide biosynthesis and metabolism
18 Protein secretion/export apparatus
19 Putative enzymes
20 Related to phage, transposon, or plasmid
21 Secreted Factors (toxins, enzymes, alginate)
22 Transcription, RNA processing and degradation
23 Transcriptional regulators
24 Translation, post-translational modification, degradation
25 Transport of small molecules
26 Two-component regulatory systems

Up-regulated genes
Down-regulated genes

Figure S3-1. Characterization of Archetype 4. Number of up- and down-regulated genes within 26 gene ontology classes for archetype 4. Enriched gene-ontology classes are high-lighted in green and red for up- and down-regulated genes respectively.

Chapter 6

Paper 2

Systems-based analysis of metabolic evolution during pathogen adaptation to the human host

Thøgersen J. C., Bartell J. A., Thykaer J., Nielsen K. F., Johansen H. K., Papin J. A., Molin S., Jelsbak L.

Manuscript submitted for publication

TITLE

Systems-based analysis of metabolic evolution during pathogen adaptation to the human host

Running title: Metabolic rewiring in adapting human pathogen

AUTHORS

Juliane C Thoegersen[1]†, Jennifer A Bartell[2]†, Jette Thykaer[1], Kristian F Nielsen[1], Helle K Johansen[3], Jason A Papin[2]*, Soeren Molin[1, 4], Lars Jelsbak[1]*

AFFILIATIONS

[1]Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[2]Biomedical Engineering, University of Virginia, Charlottesville, 22908 Virginia, USA

[3]Department of Clinical Microbiology 9301, Rigshospitalet (Copenhagen University Hospital), 2100 Copenhagen Ø, Denmark

[4]Center for Biosustainability, Technical University of Denmark, 2970 Hørsholm, Denmark

† These authors contributed equally to this article.

***Correspondence:**

Lars Jelsbak

Department of Systems Biology, DTU

Building 301, DK-2800 Kgs. Lyngby, Denmark

Phone: (+45)45256129

Email: lj@bio.dtu.dk

Jason Papin

Department of Biomedical Engineering, UVA

MR5 2041, 415 Lane Rd, Charlottesville, VA 22903, USA

Phone: (+1)4349248195

Email: papin@virginia.edu

ABSTRACT WORD COUNT: 237

TEXT WORD COUNT: 6856

ABSTRACT

Successful bacterial pathogens must satisfy specific metabolic requirements to avoid eradication by the human host during chronic infections. Identification of metabolic pathways that adapt during the course of an infection provides novel targets for potential therapeutic intervention, but distilling broad metabolic changes into specific and targetable mechanistic contributors to adaptation is challenging. We use long-term *Pseudomonas aeruginosa* infections in cystic fibrosis patients as a model system to study this evolutionary process and interrogate adapting metabolic pathways using an integrated computational and experimental systems approach. Activity in central metabolism of patient isolates was determined experimentally using growth profiling and isotope-labeling experiments. We developed isolate-specific genome-scale metabolic models through integration of transcriptomic and genomic data; these models contextualize our experimental findings from a systems perspective and elucidate specific and novel pathway adaptations during chronic infections in the CF lung. We find strong experimental evidence for a shift in metabolism towards fixation of carbon dioxide through reversal of the glycine cleavage system, which may operate as an alternative redox recycling reaction as supported by our computational modeling. This particular metabolic shift may be necessary for the bacteria to survive the oxidative stresses in the human lung environment; we provide support for this hypothesis with computational predictions of isolate-specific essential genes and altered redox pathway activity. Redox-related metabolic adaptation merits greater consideration as an important enabler of pathogen persistence and a potential therapeutic target in *Pseudomonas aeruginosa* and other emerging pathogens.

IMPORTANCE

While studies of chronic microbial infections have identified virulence factors that contribute to initial pathogen colonization as well as mutated transcriptional regulators that appear to alter a broad network of pathways, there is more limited insight into mechanistic metabolic changes that also directly contribute to successful pathogenesis. Here we study the metabolic adaptation of cystic fibrosis isolates

of *Pseudomonas aeruginosa*, a highly problematic nosocomial pathogen known for its resistance to treatment. We use an integrated framework of systems modeling and experiments to perform a comprehensive examination of changes in metabolic activity. Our results highlight an array of broad systemic changes, identify targetable genes that are essential for these changes, and propose important adaptive changes in redox metabolism through unorthodox activity of the glycine cleavage system as provider of a novel route of carbon fixation.

INTRODUCTION

Opportunistic pathogens change their metabolism in response to the conditions they encounter when they colonize their host. This metabolic reprogramming is facilitated through complex regulatory and metabolic networks encoded in the genomes of bacterial pathogens. Metabolic adaptation is necessary to capitalize on available nutrients for growth and survival and such shifts are essential for successful pathogenesis (1–3). However, the underlying metabolic mechanisms that contribute to colonization and persistence are unclear in many bacteria and how these metabolic systems develop during pathogen adaptation remains unknown. The opportunistic pathogen *Pseudomonas aeruginosa* is an ideal model system for understanding these processes. It has principally evolved in its natural habitat outside the human host, where its specific regulatory and metabolic repertoire enables growth in soil and water environments. However, during chronic infections with *P. aeruginosa* in cystic fibrosis (CF) patients, some clinical *P. aeruginosa* strains have developed into host-specific organisms by adaptive mutations that enhance survival in the human lung environment (4, 5). How these particular *P. aeruginosa* strains persist to become the dominant, chronic pathogens in the lung in contrast to other initial infecting species is poorly understood. Improving our understanding of microbial adaptation to the human host will have important implications in treatment of infectious disease, development of probiotic therapies, and other applications.

Compared to the natural environment of *P. aeruginosa*, the host environment is characterized by a complex and novel combination of stressors that could be mitigated by various adaptive strategies. In the CF lung, most of the bacterial population grows within CF sputum, which is rich in nutrient sources and has a low oxygen tension (6–8). Patient airways have elevated numbers of polymorphonuclear neutrophilic leukocytes (PMNs), alveolar macrophages and antibodies; phagocytosis of bacteria by PMNs promotes the generation of reactive oxygen species (ROS) by host cells (9). In addition to the host immune response, *P. aeruginosa* is exposed to a range of antibiotics during the course of infection in CF patients and resistance towards antibiotics is a common feature observed for adapted strains (10–12).

We know from previous studies that adaptation of *P. aeruginosa* to the CF lung environment involves many gene regulatory mutations that affect metabolism as well as mutations in metabolic enzymes (4), but connecting these underlying mechanisms to their global impacts on pathogen behaviour is difficult. The aim of this study is to identify novel metabolic systems that contribute to successful adaptation of *P. aeruginosa* in the host. Specific metabolic pathways that are undergoing changes in activity during the course of adaptation may be essential for the bacteria in order to persist in the lungs of the patients and could serve as targets for future antibiotics.

We identify these pathways of interest using a systems-level computational and experimental approach to characterize and compare the metabolic activity of clinical bacterial isolates. By integrating and contextualizing our multi-scale experimental data using a genome-scale computational metabolic model, we can streamline the prediction and comparison of early and late stage isolate phenotypes to connect shifts in the activity of a single enzyme to systems-level changes in metabolism. The results have broad implications in understanding mechanisms underlying pathogen adaptation in chronic infections and microbial evolution under selective pressures.

RESULTS

Previously, we studied genomic evolution in an epidemic *P. aeruginosa* clone type (DK2) during its dissemination across multiple patients over a 40 year time period (4, 13). The DK2 clone has been successfully transmitted between patients and replaced previously colonizing *P. aeruginosa* clone types (14). Thus, DK2 is highly adapted to the CF airway environment, which likely includes optimization of its metabolic activity for growth within the CF lung. Here, we use DK2 patient isolates to study metabolic adaptation, focusing on DK2-WT (which represents the ancestral genotype at the time of first colonization of the CF niche), and two isolates collected at later stages of clone evolution (DK2-91 and DK2-07) representing host-adapted isolates. We also included the well-studied reference strain *P. aeruginosa* PAO1 (PAO1) (see *Materials and Methods* for detailed description of the strains).

Major changes in central metabolism occur during adaptation

Growth experiments with DK2-WT, DK2-91 and DK2-07 in glucose minimal medium showed a significant reduction in growth rate for DK2-91 (μ_{max} = 0.46 h^{-1}) and DK2-07 (μ_{max} = 0.23 h^{-1}) compared to DK2-WT (μ_{max} = 0.87 h^{-1}) (Fig. S1). The growth rate of DK2-WT was higher but similar to the growth rate of PAO1 (μ_{max} = 0.63 h^{-1}). Closer inspection of the growth curves revealed diauxic growth curves for DK2-91 and DK2-07, which were not observed for PAO1 and DK2-WT. This observation led us to hypothesize that DK2-91 and DK2-07 excreted one or more metabolites that were later degraded and metabolized after glucose depletion. We then measured extracellular metabolites via GC-MS analysis, (Fig. S2) revealing that the oxidized glucose derivatives gluconate and 2-ketogluconate were accumulating in the medium for DK2-91 and DK2-07. Gluconate and 2-ketogluconate were not detected for the reference strain PAO1 and only a small amount of gluconate was detected for DK2-WT. These results indicate activity in the oxidative route of glucose degradation via gluconate and 2-ketogluconate for DK2-91 and DK2-07 as an alternative to the phosphorylative route where glucose is phosphorylated to glucose-6-phosphate (Fig. 1).

134 ***Pseudomonas* shifts its metabolism towards fixation of carbon dioxide**

135 To further investigate the central metabolism of *P. aeruginosa,* we performed substrate-labelling
136 experiments. We used a mixture of [1-^{13}C]-labelled glucose and [$^{13}C_6$]-labelled glucose. Using
137 uniformly [$^{13}C_6$]-labelled glucose has been referred to as reciprocal labelling and it is particularly useful
138 for investigating catabolism of co-substrates (15). By increasing the background labelling of position 2-5
139 of glucose, incorporation of an unlabelled carbon source (*e.g.* carbon dioxide) could be detected. The
140 [1-^{13}C]-labelled glucose can be used to track activities of different convergent pathways since the
141 labelled C-atom will end up at different positions in the carbon skeleton of metabolic intermediates
142 depending on which pathway is used to degrade glucose. This method cannot be used to differentiate
143 between the phosphorylative and oxidative route of glucose degradation to 6-phosphogluconate (Fig. 1)
144 since the resulting carbon skeleton of 6-phosphogluconate is the same regardless of the two alternative
145 routes. However, the method makes it possible to distinguish between the three glycolytic pathways: the
146 Embden-Meyerhof Parnas (EMP) pathway, the pentose phosphate (PP) pathway and the Entner-
147 Doudoroff (ED) pathway (16). Different labelling patterns of pyruvate occur depending on which
148 pathway is used to catabolize glucose. By inspecting the labelling patterns of amino acids derived from
149 pyruvate (Fig. S4), we found that the labelling degree of the carbon atom at position 1 in pyruvate was
150 around 50% for all strains grown in 100% [1-^{13}C]-glucose indicating that most if not all glucose was
151 degraded through the ED pathway. It is well known that the EMP pathway is inactive in *Pseudomonas*
152 species due to a missing enzyme and the PP pathway has previously been found only to serve
153 biosynthetic purposes for other *Pseudomonas* species including *P. putida* and *P. fluorescens* (17–20).
154
155 When we further examined the labelling patterns of amino acids derived from central metabolites from
156 the combined [1-^{13}C]-glucose and $^{13}C_6$-glucose experiment, we found that glycine had a significantly
157 lower labelling degree than the labelled carbon substrate the cells were growing on for all strains (Fig.
158 2A). This observation was most noteworthy for DK2-91 and DK2-07. The minimal medium contained

159 56% $^{13}C_6$-glucose and 44% [1-^{13}C]-glucose and we would therefore expect the average labelling degree
160 for each carbon atom to be 56% at minimum. Surprisingly, the data showed that the carbon atoms in
161 glycine had an average labelling degree of approximately 30% for DK2-91 and DK2-07 compared to
162 approximately 50% for PAO1 and DK2-WT; all were significantly lower than 56%. Since the labelled
163 substrate was the only carbon source available for the cells in the experiment, we hypothesized that the
164 cells have the capacity to fix carbon through glycine metabolism.
165
166 A literature search identified instances of non-canonical reversal of the glycine cleavage system (GCS)
167 in *Clostridia* species (21), using CO_2 as a carbon source for the synthesis of glycine. To test this, we
168 added ^{13}C-labeled bicarbonate into a growing culture in minimal medium with unlabelled glucose. Since
169 the bicarbonate was the only source of the ^{13}C isotope (except for 1.1% natural prevalence), any excess
170 labelling on glycine would indicate CO_2 fixation. Bicarbonate was added during exponential growth and
171 DK2-91 and PAO1 cells were harvested after one and two generation times. The results were a
172 qualitative measure of the ability of the cells to fix CO_2 (see *Materials and Methods*). Fig. 2B shows the
173 labelling patterns of glycine for PAO1 and DK2-91 from the ^{13}C-bicarbonate experiment. We find a
174 significant enrichment of ^{13}C in glycine for both PAO1 and DK2-91 one generation time after ^{13}C-
175 bicarbonate addition and for PAO1 the same observation was made after two generation times. Based on
176 these results we confirmed that CO_2 could be fixed into glycine when *P. aeruginosa* is growing in
177 glucose minimal medium. No significant enrichment of ^{13}C-isotope was measured after two generation
178 times in DK2-91, but since the generation time of DK2-91 is 1.4 fold longer than PAO1, dilution and
179 vaporization of ^{13}C-bicarbonate during the course of the experiment can account for this difference. We
180 included the labelling patterns of serine in Fig. 2C since serine can be produced from glycine. We find a
181 significant enrichment of the ^{13}C-isotope for DK2-91 harvested two generation times after ^{13}C-
182 bicarbonate addition. The lack of ^{13}C-labeling in serine for the other samples confirms that carbon
183 dioxide is fixed directly into glycine and not into upstream metabolic intermediates in glycolysis, since

184 otherwise we would expect at least the same degree of labelling in serine as for glycine. In conclusion,
185 we find that the *Pseudomonas* strains are incorporating carbon dioxide into glycine and under normal
186 laboratory conditions with normal carbon dioxide pressure in glucose minimal medium (Fig. 2A), this
187 observation is more pronounced in DK2-91 and DK2-07 compared to DK2-WT and PAO1. Our
188 experimental analysis of central metabolism therefore resulted in two specific findings related to
189 metabolic adaptation in the late stage isolates: (1) metabolism is shifted towards excretion of gluconate
190 and 2-ketogluconate and (2) the activity of the glycine cleavage system is altered to enable fixation of
191 carbon dioxide in a non-canonical reversal of associated pathways.
192
193 **The glycine cleavage system may operate as an alternative redox recycling reaction**
194 The ability of *P. aeruginosa* to fix carbon dioxide into glycine has not been reported previously, and
195 alterations in glycine synthesis indicated by our labelling experiments support our hypothesis that
196 carbon fixation is occurring through the glycine cleavage system (Fig. 2D) (21). A potential selective
197 advantage for the activity of this altered glycine synthesis route may be linked to the regeneration of
198 NAD^+ from NADH coupled to this reaction. We propose that this unconventional pathway phenotype
199 operates as an electron sink for the recycling of reduced electron carriers, alleviating redox stress as also
200 suggested for some anaerobic bacteria (21). Different factors in the lung environment may contribute to
201 redox stress in *P. aeruginosa* including oxidative stress from immune system defenses and low
202 availability of electron acceptors. The relative contribution of antibiotic exposure to increased oxidative
203 stress in bacteria is currently under debate (22, 23), but most recently, Dwyer et al. (24) provided
204 evidence for antibiotic-induced redox alterations in *E. coli*. We cannot specify whether the source of
205 redox stress is limited oxygen, antibiotic exposure or the immune defense within the CF lung; it is
206 possibly a combination of all three factors. However, we hypothesize that the impact of these stressors is
207 substantial, driving the enhanced carbon fixation into glycine for the late-stage clinical isolates and
208 therefore improving the balance of redox equivalents.

Metabolic models evaluate the feasibility of adapted redox metabolism

In light of our experimental observations, we focused our studies on the global effects of our proposed isolate-specific phenotypes of glycine metabolism via a well-curated and recently updated genome-scale metabolic model of *P. aeruginosa*, iPA1139 (http://bme.virginia.edu/csbl/downloads-pseudomonas-v3.php). This approach allowed us to systematically evaluate our observed phenotypes in context with model-integrated transcriptomics and sequencing data. Our experimental examination of glycine metabolism supports non-canonical activity for two connected routes of carbon fixation: glycine dehydrogenase and formate dehydrogenase (as shown in Fig. 2D). To create isolate-specific models using iPA1139, we first altered the possible activity of the glycine cleavage system and formate dehydrogenase (both canonically modelled in the forward direction); we allowed these reactions to run only in the reverse direction in our *in silico* models of DK2-91 and DK2-07 (mDK2-91 and mDK2-07) while they were modelled as reversible in our *in silico* model of DK2-WT (mDK2-WT) and the base model (iPA1139) during our data integration process. This enabled us to evaluate the feasibility and systemic impacts of the novel carbon fixation phenotypes indicated by our labelling experiments.

We additionally constrained the models to replicate isolate-specific phenotypes in our experimental conditions by integrating isolate-specific single nucleotide polymorphisms (SNP) and transcriptome data collected during growth on glucose minimal medium; this effort substantially expands our data integration approach from our earlier study of metabolic activity within CF isolates using a previous genome-scale model of *P. aeruginosa* (25). In brief, a SNP introducing a nonsense mutation in a given gene resulted in inactivation of that gene in the model; the Sorting Intolerant From Tolerant (SIFT) algorithm (26) was used to predict the functional impact of other SNPs resulting in missense mutations. These data were interpreted as "levels" of gene activity reduction (minimal, moderate, or maximal) implemented in context with transcriptome expression levels (off, potentially active, on) consistent with the gene-protein-reaction relationships to develop activity constraints for associated reactions (further

details of our constraint-based integration of SNP and transcriptome data are described in *Materials and Methods*). These methods resulted in isolate-specific models that are consistent with the substantial activity changes in pathways suggested by our experimental observations, and also enabled prediction of activity changes that were not highlighted by analysis of the *in vitro* data.

Constrained purine metabolism activity contributes to improved redox balance during adaptation

Results from the constraint-based flux modelling support the feasibility of alterations in glycine metabolism that result in novel carbon fixation; our isolate-specific models predict comparable levels of optimal biomass production regardless of GCS and formate dehydrogenase directionality. We hypothesized that the experimental phenotypes shown by the late stage isolates might indicate a shift from aerobic growth with high biomass production to microaerobic conditions where redox cofactor recycling was prioritized in addition to biomass production. Given our additional experimental evidence for a novel route of CO_2 fixation, we evaluated the effects of limitations of O_2 and CO_2 uptake and biomass production levels on the ability for each strain model to produce redox cofactors. We specifically compared the ratio of maximized NADH vs. NAD^+ production fluxes under varied uptake constraints and growth requirements for mDK2-WT and mDK2-07, as shown in Fig. 3A. mDK2-07 predicts a stable redox cofactor production ratio across varied O2 and CO2 uptake conditions while the redox ratio of mDK2-WT varies with O_2 uptake, CO_2 fixation, and biomass production.

To identify contributors to these differential predictions between mDK2-WT and mDK2-07, we modulated the gene and SNP-based constraints applied to each model. We identified the restriction of the purine metabolism enzyme phosphoribosylformylglycinamidine synthase (*purL*) due to an applied SNP constraint (Dataset S1) as the main contributor to the stability of the redox ratio in mDK2-07. While mDK2-WT has several SNPs resulting in model reaction constraints including a SNP affecting GMP synthase (*guaA*) that also plays a role in purine metabolism, a SNP in *purL* is not present and thus

the associated activity of this reaction is unconstrained. We were surprised to find that a single SNP contributed so substantially to this phenotype of a stable redox ratio under varying uptake and growth constraints; further investigation of our model identified the functional relationship between glycine metabolism and *purL* as shown in Fig. 3B, which is a non-canonical mapping of pathways that usually would not be obviously linked together. By incrementally increasing the constraints applied to the phosphoribosylformylglycinamidine synthase reaction due to the SNP in *purL* from unconstrained (mDK2-WT phenotype) to the moderate constraints applied in mDK2-07, we showed that the redox ratio of mDK2-07 transitions to a balanced state as *purL* activity is constrained. The graded impact on redox metabolism due to the *purL* constraint is clear in microaerobic conditions at low levels of CO2 uptake; high biomass production requirements magnify the impact of *purL* constraints on the transition to a balanced redox state. We propose that the close connections between the altered glycine metabolism reactions and *purL* as shown in Fig. 3B support the potential role of *purL* as a modulator of redox recycling via reversal of the glycine cleavage pathway. Our original constraints based on the SIFT predictions of SNP impact in *purL* were a broad estimate of how function might be altered; further fine-tuning may reflect the actual degree of impact of the SNP in connection to the experimentally-observed phenotype. Ultimately, our models predict that the *purL* SNP is tightly tied to improved redox balance via novel CO_2 fixation in the late stage isolates.

Many metabolic systems may contribute to the redox balance of a cell in addition to the contributions we have shown from glycine and purine metabolism. Using flux variability analysis (FVA) (27) we evaluated potential changes in redox metabolism by comparing changes in reaction activity within reactions where redox cofactors (here defined as NAD^+, NADH, $NADP^+$, NADPH, FAD^+, and/or FADH) participate versus changes in reaction activity across all reactions as shown in "FVA activity analysis" in Dataset S2. We then used a global metric of total flux activity (the sum of the ranges between minimum and maximum potential flux predicted for all reactions using FVA in a given model

divided by the same calculation performed for iPA1139) for each isolate model normalized by the same measure in iPA1139. Albeit a coarse representation of "metabolic capability" of the network, this metric provides a single snapshot of changes in metabolism as a function of changes in the underlying network characteristics. The late stage models predict 73.2% and 74.7% of the iPA1139 global activity metric compared to 69% by mDK2-WT, indicating a total flux activity increase in mDK2-91 and mDK2-07 compared to mDK2-WT with this global metabolic metric. However, the late stage models predict 85% and 84.4% of the iPA1139 redox activity metric compared to 90.1% by mDK2-WT within the subset of reactions that utilize redox cofactors, showing a reduction in the redox-related flux activity of the late stage strain models compared to mDK2-WT. We interpret these opposing changes between global and redox metabolism potential activity as an indication of systemic shifts in redox-related reaction activity between the wild-type and late stage isolates.

Genome-scale metabolic modelling contextualizes global metabolic changes

The isolate-specific metabolic models allow us to evaluate altered activity across a far greater expanse of metabolic systems than just glucose and glycine metabolism. They account for the effects of other SNPs in addition to the *purL* SNP that we previously highlighted as well as the reprogramming of the transcriptome due to adapted regulation and/or mutation. We can readily perturb specific genes and reactions computationally to investigate both the underlying drivers and potential consequences of genetic and transcriptomic adaptations at the genome scale. Here, we performed routine predictions of essential genes and flux variability that are often used to identify novel treatment targets by prioritizing genes and reactions important for growth (28). Our results indicate broad systemic rewiring in the late stage isolates that both complement our conclusions about glycine and redox metabolism as well as highlight other potential therapeutic targets important to adaptation during adaptation to a host environment.

Essential metabolic activity alters during adaptation

The isolate-specific models enable us to evaluate genes essential to strain growth phenotypes in glucose minimal medium by inactivating a given gene in the models and then predicting maximum possible growth *in silico*. Fig. 4 shows a Venn diagram categorizing all essential genes across our base model iPA1139, mDK2-WT, mDK2-91, and mDK2-07 together with a stacked histogram of reactions associated with the DK2-specific essential genes. The full list of essential genes is available in Dataset S2.

Isolate-specific SNPs were located in six genes predicted to be essential for growth in all models. Of these, constraints applied due to the SNP in PA3769, encoding GMP synthase (*guaA*), were the main driver of reduced *in silico* growth in mDK2-WT compared to the base model; constraints applied due to a SNP in PA1609, encoding beta-ketoacyl-ACP synthase I (*fabB*), affected growth to a lesser degree in the same strain. In contrast, constraints based on the *purL* SNP located in PA3763 were the main driver of reduced *in silico* growth in mDK2-91 and mDK2-07. The presence of SNPs in these genes predicted to be critical in metabolic activity according to our computational models adds emphasis to their potential importance to adaptive selection during infection.

While an array of interesting pathways have altered gene essentiality between strains, we found the changes in glucose metabolism, glycine metabolism, and oxidative phosphorylation as indicated in Fig. 4B to be of particular interest when compared with the results of our previous experiments. These changes are the result of integrating SNP and expression data into our models; we therefore can identify the experimental data underlying the specific constraints that contribute to these gains in gene essentiality. Upregulated pentose phosphate pathway genes in DK2-91 and DK2-07 contribute to differences in essentiality predicted by the late stage isolate models, highlighting adaptation in glucose metabolism. Select glycine cleavage system genes are essential in mDK2-07 due to expression-based

334 constraints; glycine dehydrogenase is identified as an essential reaction in mDK2-91 and mDK2-07 in
335 contrast to mDK2-WT for similar reasons. In oxidative phosphorylation, there is a switch in preferred
336 cytochrome complexes in oxidative phosphorylation between model mDK2-WT and mDK2-91, which
337 rely on cytochrome bc1 complex genes (PA4429-4431), while model mDK2-07 relies on cytochrome c
338 oxidase genes (PA1317-1321). This phenotype results from transcriptomic changes in DK2-91 and
339 DK2-07 compared to DK2-WT in glucose minimal medium that shows significant downregulation of
340 the *nuo* operon encoding NADH dehydrogenase (complex I of the electron chain) in the late stage
341 isolates (Dataset S3). The lack of active oxidative phosphorylation could explain the need for alternative
342 redox recycling reactions such as glycine synthesis through the glycine cleavage system. These
343 hypotheses regarding mechanistic drivers of altered essentiality between strains are a key contribution
344 enabled by our integrated systems approach. Identifying the strain-specific genes important to the
345 adaptations occurring in the DK2 lineage allows us to highlight functionally impactful SNPs and offer
346 specific, novel treatment targets within key pathways reprogrammed during evolution within the host.
347
348 **Changes in pathway activity highlight adapting systems**
349 We evaluated the results of flux variability analysis that predicts the minimum and maximum levels of a
350 reaction's flux while maintaining maximum biomass production; this enables calculation of the range of
351 potential activity for a given reaction. Fig. 5 shows a full-scale map of the metabolic network where
352 directional differences in adapting reaction activity in mDK2-WT and mDK2-07 are identified by
353 reaction colour and dashed lines identify SNPs in associated reactions. Decreases in the range of
354 reaction activity likely indicate a SNP- or gene expression-associated constraint, while increases in
355 range could be interpreted as increased flexibility of this pathway that is required to enable the
356 expression-associated constraints or produce necessary biomass components by an alternate pathway;
357 broadly, altered range in either direction may indicate areas of potentially important metabolic
358 adaptation.

Notable trends visualized on the map include increased constraint of *"Purine metabolism"* flexibility in mDK2-07 and changes in range of reaction activity in *"Glycine, serine & threonine metabolism"*. These specific metabolic pathways were also identified through our study of central metabolism. However, the network map includes a list of additional metabolic pathways with differential activity including pathways related to *"Lysine degradation"*, *"Folate metabolism"*, *"Valine, leucine, and isoleucine degradation"*, *"Pyrimidine metabolism"*, and *"Histidine metabolism"*. The mentioned pathways showed the highest degree of altered system activity in comparisons between early and late stage isolates (see *Materials and Methods*). Our systems analysis highlights areas of potential adaptation due to SNPs and altered transcriptomics in a broad array of pathways, suggesting new avenues of future experimental investigation that could elucidate other important mechanisms of adaptation in addition to our novel relationship between altered carbon fixation and redox metabolism.

DISCUSSION

In this study we have used a systems biology approach to investigate metabolic behaviour during adaptation of a pathogen to the human host. We used genome-scale metabolic models integrated with high throughput data to evaluate the feasibility and potential impacts of novel metabolic adaptations suggested by experimental characterization of glucose metabolism in *P. aeruginosa* clinical isolates. There is value in evaluating both broad changes in high level systems and specific, detailed molecular mechanisms using systems biology approaches; the former enables the prediction of systemic network production and quantification of network elements while the latter offers specific hypotheses regarding functional roles of the smallest network components (29). Here, we provide a systems level perspective of key pathways connected to metabolic adaptation, but focus our analysis on specific systems suggested by targeted experiments that indicated major changes in metabolism between initial infecting strains of *P. aeruginosa* and late-stage clinical isolates. We confirmed that the ED pathway is the only active glycolytic pathway in *P. aeruginosa,* consistent with other *Pseudomonas* species. Experimental profiling

identified a transition towards accumulation of gluconate and 2-ketogluconate and enhanced fixation of carbon dioxide into glycine specifically in the late stage isolates. Computational modelling supported the feasibility of reversed utilization of the glycine cleavage system, enabling a novel route of carbon fixation that in combination with a previously inconspicuous mutation in purine metabolism contributed to improved redox balance in the late stage isolates. We identify genes and pathways key to the adaptive processes we see in the DK2 lineage using gene essentiality and flux variability analysis, which may contribute to the design of novel treatment strategies. Our approach results in a metabolic map that provides mechanistic insight into how SNP and transcriptional changes affect metabolism at a genome scale, bridging the difficult gap between molecular mechanisms and broad, system-wide adaptation and prioritizing novel areas of metabolic reprogramming that can be targeted therapeutically.

The production of gluconate has previously been observed for clinical isolates of *P. aeruginosa.* Behrends et al. (30) found that gluconate excretion is associated with higher tolerance towards antibiotics and another study by Galet et al. (31) found that gluconate produced by *P. aeruginosa* inhibits production of an antibiotic in *Streptomyces coelicolor.* In the context of the above analysis indicating the shift in redox balancing, it might also be possible that the accumulation of gluconate and 2-ketogluconate is driven by the production of two equivalents of NADPH coupled to the oxidation reactions of glucose to 2-ketogluconate via gluconate in the periplasmic space (Fig. 1). This suggestion is not necessarily in disagreement with the correlation between gluconate and antibiotic resistance since there may also be a link between NADPH generation and antibiotic resistance given the literature on antibiotics and oxidative stress (32, 33). The identification of the ED pathway as the only active glycolytic route in *P. aeruginosa* can also be linked to generation of NADPH. The ED pathway is found to be essential for glucose metabolism in other *Pseudomonas* species; in *P. putida,* its activity has recently been associated with resistance towards oxidative stress (20). The activity of these pathways

can thereby be a mechanism to accommodate the conditions within the lung environment including both antibiotics and ROS generated by PMNs, both of which are sources of oxidative stress.

The genome scale models support the potential for novel carbon fixation routes in the late stage isolates; they also enable us to connect the late stage isolate glycine metabolism phenotype and altered redox balance in microaerobic conditions to a specific SNP in purine metabolism through network analysis. A study by Ryssel *et al.* (34) recently identified upregulated purine metabolism activity as a contributor to poor stress response in *Lactococcus lactis*, citing the production of guanine nucleotides in inducing stress sensitivity (34) while a prior study had noted the essentiality of purine synthesis in *Escherichia coli* during blood infections (35). To our knowledge, adaptation in purine metabolism has not been identified as noteworthy in cystic fibrosis infections; we evaluated published genotyping studies of cystic fibrosis isolates and identified purine SNPs in other clinical isolates (36). We suggest that altered purine metabolism may be tied to the reversal of the glycine cleavage system and contributes to resultant altered redox physiology. Whether the *purL* SNP also contributes to the need for glycine production via the GCS or is a simple way to modulate effects of the reversed GCS phenotype is currently uncertain. However, it is likely that the downregulated oxidative phosphorylation highlighted by our late stage isolate gene essentiality predictions is a way to avoid generation of oxygen radicals through the electron chain. The bacteria therefore need to redirect the metabolic flux through the glycine cleavage system to ensure regeneration of NAD^+ that is used in glycolysis.

Our hypothesis regarding the role of the glycine cleavage system as an important mediator of successful adaptation in *P. aeruginosa* led to our investigation of other cases where the glycine cleavage system is important. The glycine cleavage system is not only present in bacteria but is present across all domains of life (21). In cancer cells, elevated activity of the glycine cleavage system has been associated with tumorigenesis; glycine decarboxylase activity was correlated with reduced survival of patients with lung

cancer (37). Further investigation of the GCS in other bacterial pathogens and disordered human cells such as cancer cells may merit evaluation of a potential reversal of the pathway that enables beneficial adaptation in redox metabolism during cell proliferation in stressful environments.

Changes in metabolism in *P. aeruginosa* during adaptation have previously been considered as pleiotropic effects of regulatory actions on other targets, such as virulence factor production (38, 39). Here, we suggest that changes in metabolism are a direct target of adaptation and a driving force is selection for improved redox balance. Our systems-based analysis highlights important genes and metabolic activities involved in these adaptive processes, proposing specific pathways for novel therapeutic measures that could be used to pre-emptively combat an organism's evolutionary goals such as rewired redox metabolism. We suggest a concrete example of redox balancing through the glycine cleavage system, identifying a future target of interest for unwanted cell growth in the human body.

MATERIALS AND METHODS

Pseudomonas aeruginosa strains used in this study

We selected three isolates of the DK2 clone type for our analysis. Two of them, DK2-91 and DK2-07, are late-stage clinical isolates isolated from the same patient in 1991 and 2007, respectively (DK2-91 and DK2-07 are referred to as CF333-1991 and CF333-2007 respectively in (14)). The third isolate, DK2-WT (referred to as CF510-2006 in (13)) also shares the DK2 clone type, but has a phenotype similar to strains isolated from outside the CF lung (*P. aeruginosa* PAO1 (40) and *P. aeruginosa* PA14 (41)) and its genotype is similar to the predicted most common recent ancestor for DK2 dated back to 1970 (13). DK2-WT therefore resembles a non-adapted isolate of DK2 and this isolate serves as our point of reference for the DK2 lineage. Other early isolates of the DK2-lineage collected in the early 1970s exist. However, we chose DK2-WT as our reference for the DK2 lineage since its phylogenetic

458 branching from the most common recent ancestor is distinct from the adaptation path of DK2-91 and
459 DK2-07 in contrast to the other early isolates. We therefore expect to capture most adaptive events in
460 DK2-91 and DK2-07 by comparing to DK2-WT. *P. aeruginosa* PAO1 (PAO1) is also included as
461 reference throughout most of our experimental and *in silico* analyses. PAO1 was originally isolated from
462 a burn wound (40) and has been widely used as a reference strain for studies of *P. aeruginosa*.
463
464 **Cell storage and cultivation**
465 Cells were stored at -80°C in a 20% glycerol solution. DK2-91 and DK2-07 were streaked on a Luria-
466 Bertani (LB) agar plate and incubated at 37°C for 48 hours. Individual colonies were inoculated in 10
467 mL of morpholinepropanesulfonic acid (MOPS)-buffered medium supplemented with glucose and
468 grown aerobically at 37°C for 24-36 hours (depending on growth rate). The total composition of the
469 MOPS minimal medium was 40 mM MOPS, 9.5 mM NH_4Cl, 0.28 mM K_2SO_4, 1.3 mM KH_2PO_4, 10
470 mM glucose and vitamins (0.4 µM biotin, 10 µM pyroxidal-HCl, 2.3 µM folic acid, 2.6 µM riboflavin, 8
471 µM niacinamide, 3 µM thiamine-HCl and 2 µM pantotheneate) (42).
472 DK2-WT and PA01 were streaked on LB agar plates and incubated at 37°C for 24 hours. Individual
473 colonies were inoculated in 10 mL of MOPS minimal medium supplemented with glucose and grown
474 aerobically at 37°C for 16 hours. After initial incubation cells were transferred to a 250 mL baffled flask
475 with 50 mL MOPS minimal medium supplemented with 10 mM of defined carbon source to an optical
476 density (OD_{600}) of 0.01 measured at 600 nm.
477 Cell growth was determined by measuring OD_{600} during growth. Cells were harvested for GC-MS and
478 DNA microarray analyses at OD_{600} = 0.4 during the exponential growth phase. Supernatant was
479 collected for a glucose determination assay during growth in order to make biomass yield calculations.
480
481
482

Biomass yield calculations

Glucose concentrations were determined enzymatically using a glucose reagent (catalogue no. 7200-017A, from Thermo Electron, Australia). The dry weight biomass concentration was estimated using a correlation factor of 0.360 g cellular dry weight per OD unit. This correlation factor was determined for an *Escherichia coli* strain (43) and is assumed to be valid for *P. aeruginosa.* The biomass yield on glucose was determined using the concentration data for biomass and glucose, respectively.

Labelling experiments

The experimental protocol for labelling determination was modified from (44). Cells were grown in MOPS minimal medium to an OD_{600} of 0.4. 10 mM [1-^{13}C]-glucose (D-glucose-^{13}C, 99% ^{13}C, from Isotec, Miamisburg, Ohio, USA, CAS no. 297046) was used as a carbon source. For some experiments, a mixture of 44 mol-% [1-^{13}C]-glucose and 56 mol-% $^{13}C_6$ glucose (D-glucose-$^{13}C_6$, 99 % ^{13}C, from Isotec, Miamisburg, Ohio, USA, CAS no. 110187-42-3) were used to give a final glucose concentration of 10 mM.

30 ml culture was harvested and the samples were spun down for 10 minutes at 5,000 rpm at 4°C. The pellet was resuspended in 2 mL 0.9 % NaCl and the volume was divided into two Eppendorf tubes. The Eppendorf tubes were further spun down for two minutes at 10,000 rpm at 4°C and the pellets were finally stored at -80°C until hydrolysis and subsequent derivatization and amino acid analysis by GC-MS. The supernatant was stored in 4 individual Eppendorf tubes of 1 mL at -80°C for later GC-MS analysis of extracellular metabolites. Proteinogenic amino acid analysis from ^{13}C-labeled biomass and GC-MS analysis for extracellular are fully described in Supplemental Text S1.

Testing for CO_2 incorporation into glycine

PAO1 and DK2-91 were grown in MOPS minimal medium supplemented with 10 mM *unlabelled* glucose. At OD_{600}=0.01 20 mM of $NaH^{13}CO_3$ was added. Cells were harvested at OD_{600}=0.2 and

OD_{600}=0.4 and labelling patterns of amino acids were determined as described above. We chose PAO1 instead of DK2-WT to find out if the potential of carbon fixation into glycine is general for *P. aeruginosa* or just a feature of the DK2 lineage. We chose DK2-91 to represent the late-stage clinical isolates, since the growth rate of DK2-91 was higher than that for DK2-07 (slow growth of the cells would allow more bicarbonate to vaporize before cell harvest). We used a high concentration of bicarbonate (20 mM) to make sure that some bicarbonate would remain in the medium at the time of harvesting despite dilution with unlabelled bicarbonate/carbon dioxide and vaporization. Ideally, the experiment would be carried out with concentrations of bicarbonate and carbon dioxide corresponding to the initial experiments. However, the labelled bicarbonate and carbon dioxide would be diluted out from unlabelled carbon dioxide produced under glycolysis in the growing culture. Therefore this experimental setup only addresses the question whether carbon dioxide is fixated in glycine synthesis and the results cannot be used quantitatively.

DNA microarray analysis

Cells were grown in MOPS minimal medium supplemented with 10 mM glucose to an OD_{600} of 0.4 prior to Affymetrix *P. aeruginosa* GeneChip microarray analysis. Microarray data were generated using Affymetrix protocols as previously described (4). Data processing was carried out according to Thøgersen et al, 2013 (46). The raw *cel*-files were extracted in R by use of the package *affy* (47) followed by qspline normalization (48) and calculation of gene expression index values using *robust multiarray average* expression measure (49). Differentially expressed genes for DK2-91 and DK2-07 compared to DK2-WT were determined with Bonferroni adjusted p-values (significance level p=0.05) using the R package "*limma*" (50). Enriched gene ontology classes among differentially expressed genes were identified by the Hypergeometric distribution test with significance level p = 0.01.

Sorting Intolerant from Tolerant (SIFT) Analysis of SNPs

The SNP data were obtained from previous studies of the DK2-WT, DK2-91 and DK2-07 strains (4, 13). Strain specific SNPs are listed in Dataset S1 including our SIFT (Sorting Intolerant From Tolerant (26)) analysis of missense mutations in metabolic genes. The SIFT analysis is used to predict if a missense mutation would affect protein function of the given gene product, providing numeric scores that indicate the degree to which a missense SNP is tolerated or affects protein function.

Isolate-specific genome-scale metabolic models

The genome scale metabolic reconstruction for *P. aeruginosa* PAO1, iPA1139, was used as the base for all computational modelling in this study. This model accounts for the function of 1139 genes, 1491 reactions, and 1280 metabolites involved in the metabolism of *P. aeruginosa*. Isolate-specific genome-scale metabolic models were created by a semi-automated approach in order to incorporate both SNP and gene expression-based constraints using the TIGER Toolbox 1.2.0 (51, 52). Further details on construction of the isolate specific models are included in Supplemental text S1, the resulting isolate specific models are included in Dataset S4, and base model iPA1139 and SBML versions of the isolate-specific models are available at http://bme.virginia.edu/csbl/downloads-pseudomonas-v3.php.

The isolate specific models were then used to evaluate metabolic activity using several methods of constraint-based modelling. Flux balance analysis (FBA) was used to predict the ability of each isolate model to grow (produce biomass) in 'wild type' conditions as well as with single genes deleted to identify *in silico* genes essential for growth. Flux variability analysis (FVA) was used to predict changes in potential reaction activity by calculating the minimum and maximum flux of a given reaction when the model was required to produce maximum biomass. We calculated the flux range from the maximum and minimum flux values for each reaction, and then determined whether the range increased or decreased compared to unconstrained iPA1139, sorting results into 5 categories: decreased range in

mDK2-07 compared to mDK2-WT, increased range in mDK2-07 compared to mDK2-WT, or comparable changes in the ranges in both strains that are increased, the same, or decreased compared to iPA1139. To identify subsystems with high concentrations of changes in activity that we interpret here as potential adaptive processes, we counted the number of reactions in these altered activity categories within each subsystem and then normalized by the number of active subsystem reactions in iPA1139 as shown in Dataset S2.

The redox cofactor production analysis presented in Fig. 3 was performed by optimizing for the maximum production capacity of NAD+ and NADH separately while constraining the maximum uptake rates of O_2 and CO_2 to 0.2, 2, 5, and 20 mmol/gDW/hr (low to high uptake rates) and fixing the production rate of biomass at 0 to 100% of optimum production when only growth is maximized. The command-line implementation of Metdraw (also available at www.metdraw.com) was used to build a full-sized map of the base model on which FVA results were overlaid automatically (53).

ACKNOWLEDGMENTS

The authors thank Alexander Rosenkjær, Jesper Mogensen, Mikkel Lindegaard and Susanne Kofoed for assistance with the experiments, and Edik Blais for helpful comments regarding the modelling analysis. Additionally, we thank Mogens Kilstrup and Rasmus Marvig for helpful discussions and constructive feedback on the manuscript.

This work was funded by the National Institutes of Health (NIH RO1 GM088244) to JAP, the Danish Council for Independent Research (10-084969) to SM, and the Villum Foundation (VKR023113) to LJ. The Novo Nordisk Foundation supported HKJ as a clinical research stipend.

583 **AUTHOR CONTRIBUTIONS:**

584 JCT, JAB, JT, JAP, SM and LJ participated in the design of the study. JCT performed the *in vitro*
585 experiments and assisted in the modelling analysis. JT and KFN guided the *in vitro* experiments. JAB
586 performed the modelling analysis. HKJ analysed clinical information. JCT and JAB conducted the
587 interpretation of results and wrote the manuscript. JAP, SM and LJ helped to draft the manuscript. All
588 authors read and approved the final manuscript.

589

590 **COMPETING FINANCIAL INTERESTS**

591 The authors declare that they have no competing interests.

592

593 **DATA AVAILABILITY**

594 The DNA microarray data discussed in this publication have been deposited in NCBI's Gene Expression
595 Omnibus (54) and are accessible through GEO Series accession number GSE62970 (private link for
596 reviewers:
597 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=enypqggkdlcdlif&acc=GSE62970).
598 The genome-scale metabolic model of *P. aeruginosa*, iPA1139, is available at the website hosting other
599 Papin lab metabolic reconstructions (http://bme.virginia.edu/csbl/downloads-pseudomonas-v3.php).

600

601 **REFERENCES**

602 1. **Brown SA, Palmer KL, Whiteley M**. 2008. Revisiting the host as a growth medium. Nat. Rev.
603 Microbiol. **6**:657–66.

604 2. **Eisenreich W, Dandekar T, Heesemann J, Goebel W**. 2010. Carbon metabolism of
605 intracellular bacterial pathogens and possible links to virulence. Nat. Rev. Microbiol. **8**:401–12.

606 3. **Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, Kupiec M, Dahan O, Pilpel Y**.
607 2009. Adaptive prediction of environmental changes by microorganisms. Nature **460**:220–4.

608 4. **Yang L, Jelsbak L, Marvig RL, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson**
609 **A, Johansen HK, Ciofu O, Høiby N, Sommer MO a, Molin S**. 2011. Evolutionary dynamics of
610 bacteria in a human host environment. Proc. Natl. Acad. Sci. U. S. A. **108**:7481–6.

611 5. **Folkesson A, Jelsbak L, Yang L, Johansen HK, Ciofu O, Høiby N, Molin S**. 2012. Adaptation
612 of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective. Nat. Rev.
613 Microbiol. **10**:841–51.

614 6. **Worlitzsch D, Tarran R, Ulrich M, Schwab U, Cekici A, Meyer KC, Birrer P, Bellon G,**
615 **Berger J, Weiss T, Botzenhart K, Yankaskas JR, Randell S, Boucher RC**. 2002. Effects of
616 reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis
617 patients. J. Clin. Invest. **109**:317–325.

618 7. **Palmer KL, Mashburn LM, Singh PK, Whiteley M, Al PET**. 2005. Cystic Fibrosis Sputum
619 Supports Growth and Cues Key Aspects of Pseudomonas aeruginosa Physiology. J. Bacteriol.
620 **187**:5267–5277.

621 8. **Ohmani DE, Chakrabarty AM**. 1982. Utilization of Human Respiratory Secretions by Mucoid
622 Pseudomonas aeruginosa of Cystic Fibrosis Origin **37**:662–669.

623 9. **Hoiby N**. 2006. P. aeruginosa in Cystic Fibrosis Patients Resists Host Defenses, Antibiotics.
624 Microbe **1**:571–577.

625 10. **Govan JRW, Nelson W**. 1993. Microbiology of cystic fibrosis lung infections: themes and
626 issues. J. R. Soc. Med. **86**:11–18.

627 11. **Burns JL, Emerson J, Stapp JR, Yim DL, Krzewinski J, Louden L, Ramsey BW, Clausen**
628 **CR**. 1998. Microbiology of Sputum from Patients at Cystic Fibrosis Centers in the United States
629 **27**:158–163.

630 12. **Döring G, Conway S, Heijerman HGM, Hodson ME, Høiby N, Smyth A, Touw DJ**. 2000.
631 Antibiotic therapy against Pseudomonas aeruginosa in cystic fibrosis : a European consensus. Eur
632 Respir J **16**:749–767.

633 13. **Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L**. 2012. Deletion and acquisition of
634 genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host
635 environment. Environ. Microbiol. **14**:2200–11.

636 14. **Jelsbak L, Johansen HK, Frost A-L, Thøgersen R, Thomsen LE, Ciofu O, Yang L,**
637 **Haagensen JAJ, Høiby N, Molin S**. 2007. Molecular epidemiology and dynamics of
638 Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients. Infect. Immun. **75**:2214–
639 24.

640 15. **Christensen B, Nielsen J**. 2002. Reciprocal 13C-labeling: a method for investigating the
641 catabolism of cosubstrates. Biotechnol. Prog. **18**:163–6.

642 16. **Gunnarsson N, Mortensen UH, Sosio M, Nielsen J**. 2004. Identification of the Entner-
643 Doudoroff pathway in an antibiotic-producing actinomycete species. Mol. Microbiol. **52**:895–
644 902.

645 17. **Del Castillo T, Ramos JL, Rodríguez-Herva JJ, Fuhrer T, Sauer U, Duque E**. 2007.
646 Convergent peripheral pathways catalyze initial glucose catabolism in Pseudomonas putida:
647 genomic and flux analysis. J. Bacteriol. **189**:5142–52.

18. **Fuhrer T, Fischer E, Sauer U**. 2005. Experimental identification and quantification of glucose metabolism in seven bacterial species. J. Bacteriol. **187**:1581–90.

19. **Ramos J-L**. 2004. Pseudomonas - Genomics Life Style and Molecular Architecture. Kluwer Acacemic/Plenum, New York.

20. **Chavarría M, Nikel PI, Pérez-Pantoja D, de Lorenzo V**. 2013. The Entner-Doudoroff pathway empowers Pseudomonas putida KT2440 with a high tolerance to oxidative stress. Environ. Microbiol. **15**:1772–85.

21. **Bar-Even A, Noor E, Milo R**. 2012. A survey of carbon fixation pathways through a quantitative lens. J. Exp. Bot. **63**:2325–42.

22. **Kohanski MA, Dwyer DJ, Collins JJ**. 2010. How antibiotics kill bacteria: from targets to networks. Nat. Rev. Microbiol. **8**:423–35.

23. **Liu Y, Imlay JA**. 2013. Cell death from antibiotics without the involvement of reactive oxygen species. Science **339**:1210–3.

24. **Dwyer DJ, Belenky PA, Yang JH, MacDonald IC, Martell JD, Takahashi N, Chan CTY, Lobritz MA, Braff D, Schwarz EG, Ye JD, Pati M, Vercruysse M, Ralifo PS, Allison KR, Khalil AS, Ting AY, Walker GC, Collins JJ**. 2014. Antibiotics induce redox-related physiological alterations as part of their lethality. Proc. Natl. Acad. Sci. U. S. A. **111**:E2100–9.

25. **Oberhardt MA, Goldberg JB, Hogardt M, Papin JA**. 2010. Metabolic network analysis of Pseudomonas aeruginosa during chronic cystic fibrosis lung infection. J. Bacteriol. **192**:5534–48.

26. **Kumar P, Henikoff S, Ng PC**. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. **4**:1073–81.

669 27. **Mahadevan R, Schilling CH**. 2003. The effects of alternate optimal solutions in constraint-based
670 genome-scale metabolic models. Metab. Eng. **5**:264–276.

671 28. **Chavali AK, D'Auria KM, Hewlett EL, Pearson RD, Papin JA**. 2012. A metabolic network
672 approach for the identification and prioritization of antimicrobial drug targets. Trends Microbiol.
673 **20**:113–23.

674 29. **Heinemann M, Sauer U**. 2010. Systems biology of microbial metabolism. Curr. Opin.
675 Microbiol. **13**:337–43.

676 30. **Behrends V, Ryall B, Zlosnik JE a, Speert DP, Bundy JG, Williams HD**. 2012. Metabolic
677 adaptations of Pseudomonas aeruginosa during cystic fibrosis chronic lung infections. Environ.
678 Microbiol.

679 31. **Galet J, Deveau A, Hôtel L, Leblond P, Frey-Klett P, Aigle B**. 2014. Gluconic acid-producing
680 Pseudomonas sp. prevent γ-actinorhodin biosynthesis by Streptomyces coelicolor A3(2). Arch.
681 Microbiol. **3**.

682 32. **Derewacz DK, Goodwin CR, McNees CR, McLean JA, Bachmann BO**. 2013. Antimicrobial
683 drug resistance affects broad changes in metabolomic phenotype in addition to secondary
684 metabolism. Proc. Natl. Acad. Sci. U. S. A. **110**:2336–41.

685 33. **Kohanski MA, Dwyer DJ, Collins JJ**. 2010. How antibiotics kill bacteria: from targets to
686 networks. Nat. Rev. Microbiol. **8**:423–35.

687 34. **Ryssel M, Hviid A-MM, Dawish MS, Haaber J, Hammer K, Martinussen J, Kilstrup M**.
688 2014. Multi-stress resistance in Lactococcus lactis is actually escape from purine induced stress
689 sensitivity. Microbiology.

690 35. **Samant S, Lee H, Ghassemi M, Chen J, Cook JL, Mankin AS, Neyfakh AA**. 2008.
691 Nucleotide biosynthesis is critical for growth of bacteria in human blood. PLoS Pathog. **4**:e37.

692 36. **Bezuidt OK, Klockgether J, Elsen S, Attree I, Davenport CF, Tümmler B**. 2013. Intraclonal
693 genome diversity of Pseudomonas aeruginosa clones CHA and TB. BMC Genomics **14**:416.

694 37. **Zhang WCC, Shyh-Chang N, Yang H, Rai A, Umashankar S, Ma S, Soh BSS, Sun LLL, Tai**
695 **BCC, Nga MEE, Bhakoo KKK, Jayapal SRR, Nichane M, Yu Q, Ahmed DAA, Tan C, Sing**
696 **WPP, Tam J, Thirugananam A, Noghabi MSS, Huei Pang Y, Ang HSS, Mitchell W, Robson**
697 **P, Kaldis P, Soo RAA, Swarup S, Lim EHH, Lim B, Pang YH**. 2012. Glycine Decarboxylase
698 Activity Drives Non-Small Cell Lung Cancer Tumor-Initiating Cells and Tumorigenesis. Cell
699 **148**:259–72.

700 38. **Nguyen D, Singh PK**. 2006. Evolving stealth: genetic adaptation of Pseudomonas aeruginosa
701 during cystic fibrosis infections. Proc. Natl. Acad. Sci. U. S. A. **103**:8305–6.

702 39. **Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, Miller SI,**
703 **Ramsey BW, Speert DP, Moskowitz SM, Burns JL, Kaul R, Olson M V**. 2006. Genetic
704 adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. Proc. Natl. Acad.
705 Sci. U. S. A. **103**:8487–92.

706 40. **Holloway BW, Krishnapillai V, Morgan AF**. 1979. Chromosomal genetics of Pseudomonas.
707 Microbiol. Rev. **43**:73–102.

708 41. **Rahme LG, Stevens EJ, Wolfort SF, Shao J, Tompkins RG, Ausubel FM**. 1995. Common
709 virulence factors for bacterial pathogenicity in plants and animals. Science **268**:1899–902.

710 42. **Jensen PR, Hammer K**. 1993. Minimal requirements for exponential growth of Lactococcus
711 lactis. Appl. Environ. Microbiol. **59**:4363–4366.

712 43. **Kiviharju K, Salonen K, Moilanen U, Meskanen E, Leisola M, Eerikäinen T**. 2007. On-line
713 biomass measurements in bioreactor cultivations : comparison study of two on-line probes. J Ind
714 Microbiol Biotechnol. **34**:561–566.

715 44. **Christensen B, Nielsen J**. 1999. Isotopomer Analysis Using GC - MS **290**:282–290.

716 45. **Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, Fiehn O**. 2010. FiehnLib –
717 mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-
718 flight gaschromatography/mass spectrometry. Anal. Chem. **81**:10038–10048.

719 46. **Thøgersen JC, Mørup M, Damkiær S, Molin S, Jelsbak L**. 2013. Archetypal analysis of
720 diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways.
721 BMC Bioinformatics **14**:279.

722 47. **Gautier L, Cope L, Bolstad BM, Irizarry RA**. 2004. affy--analysis of Affymetrix GeneChip
723 data at the probe level. Bioinformatics **20**:307–15.

724 48. **Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild H, Brunak S,**
725 **Knudsen S**. 2002. A new non-linear normalization method for reducing variability in DNA
726 microarray experiments. Genome Biol. **3**:1–16.

727 49. **Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP**.
728 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe
729 level data. Biostatistics **4**:249–64.

730 50. **Smyth G**. 2005. Limma: linear models for microarray data, p. 397–420. *In* Gentleman, R, Carey,
731 V, Dodoit, S, Irizarry, R, Huber, W (eds.), Bioinformatics and Computational Biology Solutions
732 Using R and Bioconducter. Springer, New York.

733 51. **Jensen PA, Lutz KA, Papin JA**. 2011. TIGER: Toolbox for integrating genome-scale metabolic
734 models, expression data, and transcriptional regulatory networks. BMC Syst. Biol. **5**:147.

735 52. **Zur H, Ruppin E, Shlomi T**. 2010. iMAT: an integrative metabolic analysis tool. Bioinformatics
736 **26**:3140–2.

737 53. **Jensen PA, Papin JA**. 2014. MetDraw: automated visualization of genome-scale metabolic
738 network reconstructions and high-throughput data. Bioinformatics **30**:1327–1328.

739 54. **Edgar R, Domrachev M, Lash AE**. 2002. Gene Expression Omnibus : NCBI gene expression
740 and hybridization array data repository **30**:207–210.

741 55. **Cuskey SM, Wolff JA, Phibbs P V, Olsen RH**. 1985. Cloning of Genes Specifying
742 Carbohydrate Catabolism in Pseudomonas aeruginosa and Pseudomonas putida **162**:865–871.

743 56. **Temple L, Cuskey SM, Perkins RE, Bass RC, Morales NM, Christie GE, Olsen RH, Phibbs**
744 **P V**. 1990. Analysis of Cloned Structural and Regulatory Genes for Carbohydrate Utilization in
745 Pseudomonas aeruginosa PAOt **172**:6396–6402.

746 57. **Hunt JC, Phibbs P V**. 1983. Regulation of alternate peripheral pathways of glucose catabolism
747 during aerobic and anaerobic growth of Pseudomonas aeruginosa. J. Bacteriol. **154**:793–802.

748 58. **Conway T**. 1992. The Entner-Doudoroff pathway: history, physiology and molecular biology.
749 FEMS Microbiol. Rev. **9**:1–27.

750 59. **Winsor GL, Lam DKW, Fleming L, Lo R, Whiteside MD, Yu NY, Hancock REW,**
751 **Brinkman FSL**. 2011. Pseudomonas Genome Database: improved comparative analysis and
752 population genomics capability for Pseudomonas genomes. Nucleic Acids Res. **39**:D596–600.

753

754

755 **FIGURE LEGENDS**

756

757 **Figure 1 - Glucose metabolism in *Pseudomonas aeruginosa*.**

758 Glucose can enter the cell through the phosphorylative or the oxidative route. The oxidative route
759 involves conversion of glucose into gluconate or 2-ketogluconate. In the cytoplasm, further degradation
760 of pyruvate may occur through three alternate pathways. The blue arrows indicate alternative convergent
761 pathways and their respective names (17–19, 55–58). Fig. S3 shows an expanded version of Fig. 1
762 including gene locus names assigned to each reaction. Abbreviations: ED (Entner-Doudoroff), PP
763 (pentose phosphate), EMP (Embden-Meyerhof-Parnas), TCA (Tricarboxylic acid), P (phosphate).

764

765 **Figure 2 - Carbon dioxide fixation into glycine through the glycine cleavage system.**

766 **A** Labelling of glycine derived from cultivation in glucose minimal medium (56% $^{13}C_6$-glucose and 44%
767 [1-^{13}C]-glucose). The amount of background labelling from $^{13}C_6$-glucose (56%) is indicated as a
768 separate column.

769 **B** Labelling of glycine derived from cultivation in unlabelled glucose and labelled bicarbonate
770 ($H^{13}CO_3^-$).

771 **C** Labelling of serine derived from cultivation in unlabelled glucose and labelled bicarbonate ($H^{13}CO_3$).

772 **D** The glycine cleavage system in reverse. Two molecules of carbon dioxide are fixated into glycine -
773 one of them via formate formation. Figure adapted from (21). Abbreviations: Reduced electron donor
774 (AH_2), oxidized electron donor (A), tetrahydrofolate (THF), lipoyl protein (LP).

775

776 In **A-C** the (m+1)-columns indicate the percentages of compounds with one labelled C-atom, whereas
777 the (m+2)-columns indicate the percentages of compound with both carbon atoms labelled. The control

778 is a measure of the naturally occurring ^{13}C-isotope in bovine serum albumin (BSA). In **A**, (*) denotes
779 where the labelling percentages of (m+2)-labelling are significantly lower (Student's *t*-test, two-sided,
780 significance level, $p = 0.05$) than the background level from labelled glucose in the medium. In **B-C**, (*)
781 denotes where the percentages of (m+1)-labelling of strains are significantly higher (Student's *t*-test,
782 significance level, two-sided, $p = 0.05$) than the level of the naturally occurring isotope (control). Note
783 that panels **A-C** have different scales.
784
785 **Figure 3 - Redox cofactor production differences between mDK2-WT and mDK2-07 due to SNP in**
786 **purine metabolism.**
787 **A** An evaluation of the effects of altered O_2 and CO_2 uptake on the ratio of NADH production to NAD^+
788 production under a range of biomass production constraints for mDK2-WT (blue), mDK2-07 (red), and
789 mDK2-07 with reduced *purL* activity constraints (shades of purple).
790 **B** Pathway illustration of the connection between glycine metabolism and purine metabolism,
791 specifically highlighting *purL*, a gene that contains a SNP in DK2-07 that the model predicts is
792 connected to differential redox metabolism activity between strains. Abbreviations: Glycinamide
793 ribonucleotide (GAR), 5'-phosphoribosylformylglycinamidine (FGAM), lipoyl protein (LP).
794
795 **Figure 4 - Isolate-specific gene essentiality and associated functions.**
796 **A** Stacked histogram of reactions associated with DK2-specific essential genes, as shown by %
797 associated reactions within a particular KEGG subsystem. Total reactions assigned in the KEGG
798 subsystem are included in parentheses in each subsystem label. Results for the essential reaction
799 distribution across the base model and three isolate-specific models are shown in each subsystem
800 category as indicated by colours corresponding to the categories of the Venn diagram in panel B. Bolded
801 histogram labels highlight subsystems that show variation in reaction distributions between isolate
802 models.

B Venn diagram of the distribution of *in silico* essential gene predictions, highlighting the differences in unique versus shared essential genes between mDK2-WT, mDK2-91, mDK2-07 and the base model (iPA1139).

Figure 5 - Flux variability analysis displayed on global metabolic map.

Differential reaction activity ranges between mDK2-WT and mDK2-07 predicted by flux variability analysis under 100% biomass demands. Increase/decrease in flexibility was identified through comparison of mDK2-WT and mDK2-07 reaction predictions with base model iPA1139 reaction predictions. Dashed lines indicate SNPs present in DK2-WT and DK2-07. The map provides an overview of metabolic changes between DK2-WT and DK2-07, with enlarged panels of purine metabolism and glycine, serine and threonine metabolism presented to highlight the important changes identified in these subsystems. Users can zoom in to identify specific compounds and reactions connected to highlighted areas of differential activity. Associated implementation of the compounds and reactions can be found in the genome-scale models in Dataset S4.

821 **SUPPLEMENTAL LEGENDS**

822

823 **Supplemental Text S1 - Expanded materials and methods.**

824

825 **Figure S1 - Growth data.**

826 **A** Growth of the *P. aeruginosa* strains PA01, DK2-WT, DK2-91 and DK2-07 in

827 morpholinepropanesulfonic acid (MOPS)-buffered minimal medium supplemented with 10 mM [1-^{13}C]-

828 glucose. The curves show optical density measured at 600 nm.

829 **B** Specific growth rates for *P. aeruginosa* strains used in this study. All of the growth rates are

830 significantly different from each other (Student's *t*-test, two-sided, significance level $p = 0.05$).

831 **C** Biomass yields determined for *P. aeruginosa* strains used in this study.

832 For panels **B-C** the values are based on biological triplicates and error bars indicate standard deviations.

833 Abbreviations: Dry weight (DW).

834

835 **Figure S2 - GC-MS analysis of extracellular metabolites**.

836 Comparison of GC-MS profiles (derivatised by methoximation and silylation) of lyophilized broth,

837 showing a shift towards 2-ketogluconate and gluconate production in DK2-91 and DK2-07. Reference

838 standards of all glucose, gluconate, as well as 2- and 5-ketogluconate were co-analysed in the sequence.

839 Two chromatographic peaks are formed per sugar due to the derivatization process.

840

841 **Figure S3 - Expanded figure of glucose metabolism in *Pseudomonas aeruginosa.***

842 Glucose can enter the cell through the phosphorylative or the oxidative route. The oxidative route

843 involves conversion of glucose into gluconate or 2-ketogluconate..In the cytoplasm, further degradation

844 to pyruvate can happen through three alternate pathways. The blue arrows indicate alternative

845 convergent pathways and their respective names. Gene locus names according to *P. aeruginosa* PAO1

(59) are assigned to each reaction. Abbreviations: ED (Entner-Doudoroff), PP (pentose phosphate), EMP (Embden-Meyerhof-Parnas), TCA (Tricarboxylic acid), P (phosphate).

Figure S4 - Labelling patterns of pyruvate from the [1-^{13}C]-glucose experiment.

Summed fractional labelings (%) for derivatised amino acids (the number indicates the mass of the amino acid fragment for the lowest mass isotopomer). The shown amino acids are all derived from pyruvate and the corresponding carbon positions in pyruvate (PYR) are indicated in brackets. The values are based on biological triplicates and the error bars show the standard deviations.

The data from Valine144 show low standard deviations from triplicate analyses. We can therefore conclude with high degree of certainty that position 2 and 3 have average labelling degrees of around 1.1 % (corresponding to natural prevalence of ^{13}C).

There are some uncertainties in determining the labelling patterns of Alanine158 and Valine186. The labelling degree of the carbon atom at position 1 in Pyruvate can be calculated by subtracting the value of Alanine116 from Alanine158 or Valine144 from Valine186. This calculation gives a rough estimate on the labelling degree of the Pyruvate position 1 being around 50%.

Figure S5 - Converting 'omics' data to isolate-specific model constraints.

Counts of genes and SNPs binned into their respective functional categories (shown in top two tables) are manually evaluated for combined expression-SNP functional impact and then provided to the TIGER implementation of iMAT for constraint development. Resulting iMAT predictions of 'off' genes that should be inactivated and 'on' genes that should result in associated reactions carrying at least a minimum level of flux during growth are shown. After enforcing the requirement that all models must be able to produce biomass (grow), the resulting number of reactions with constrained flux activity in each isolate-specific model is presented in the last table.

Dataset S1. Gene expression data.

Differentially expressed genes in glucose minimal media for DK2-91 (sheet 2) and DK2-07 (sheet 3) compared to DK2-WT. Sheet 1 includes a hypergeometric distribution test of enriched gene ontology classes among differentially expressed genes for DK2-91 and DK2-07 compared to DK2-WT.

Dataset S2. SNP and expression constraints.

Data supporting the conversion of identified SNP and expression levels into constraints applied to genome scale models to create isolate-specific models. Includes SIFT predictions for each SNP (categorized by affected strain), comparison of SIFT predicted impact and gene expression constraints as implemented through gene-protein-reaction relationships, and final inactivated gene sets for each strain model. Further details are included in S8 and legends within sub-sheets of the file.

Dataset S3. In silico essential gene analysis and flux variability analysis.

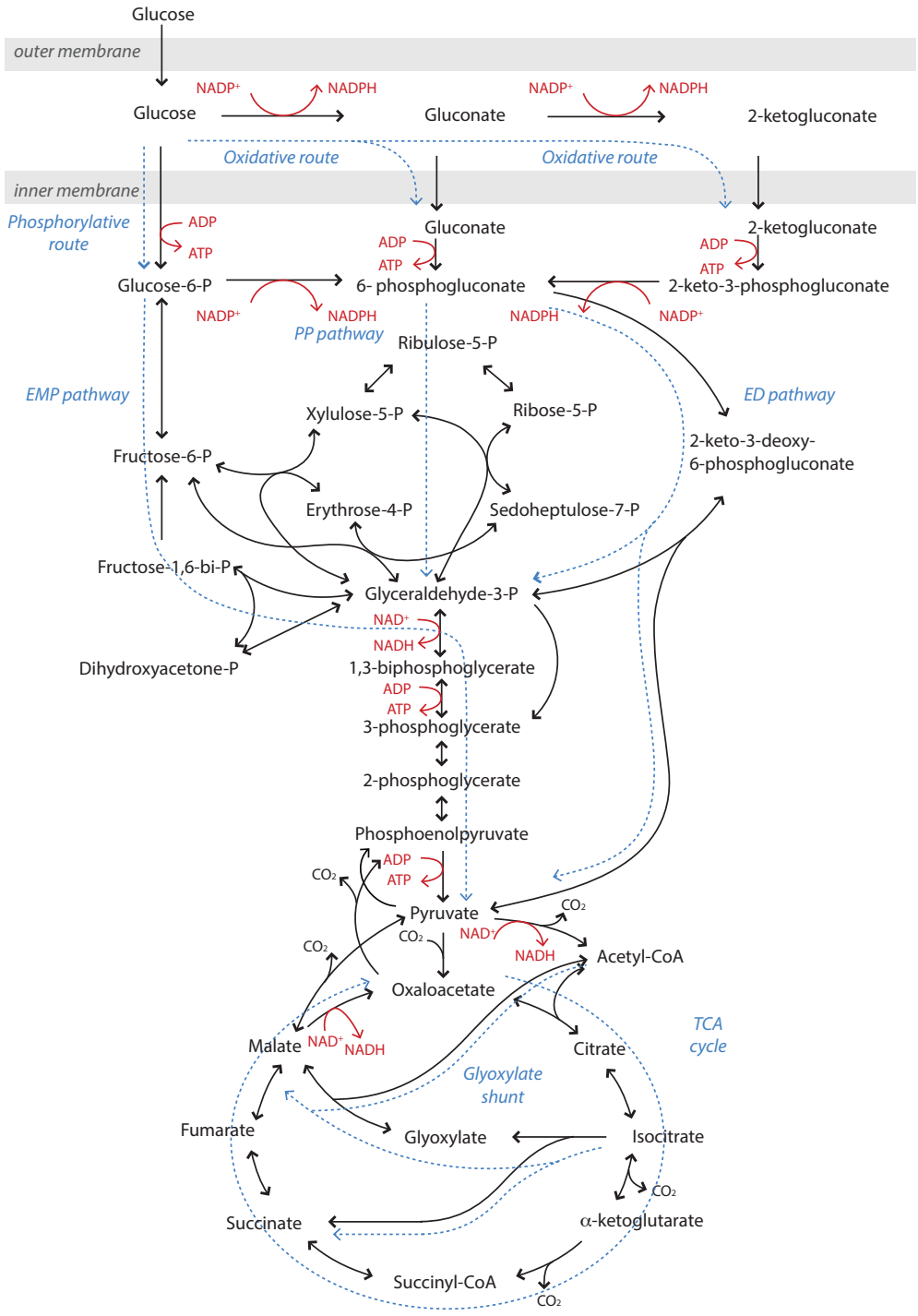
Included are essential genes predicted by iPA1139 and each isolate specific model, all reactions associated with these essential genes, flux variability analysis for all reactions under 100% biomass production constraints, subsystem-based FVA analysis, and FVA-based comparison of redox metabolism activity to global activity. Further details are included in Dataset S4 and legends within sub-sheets of the file.

Dataset S4. Genome scale metabolic models.

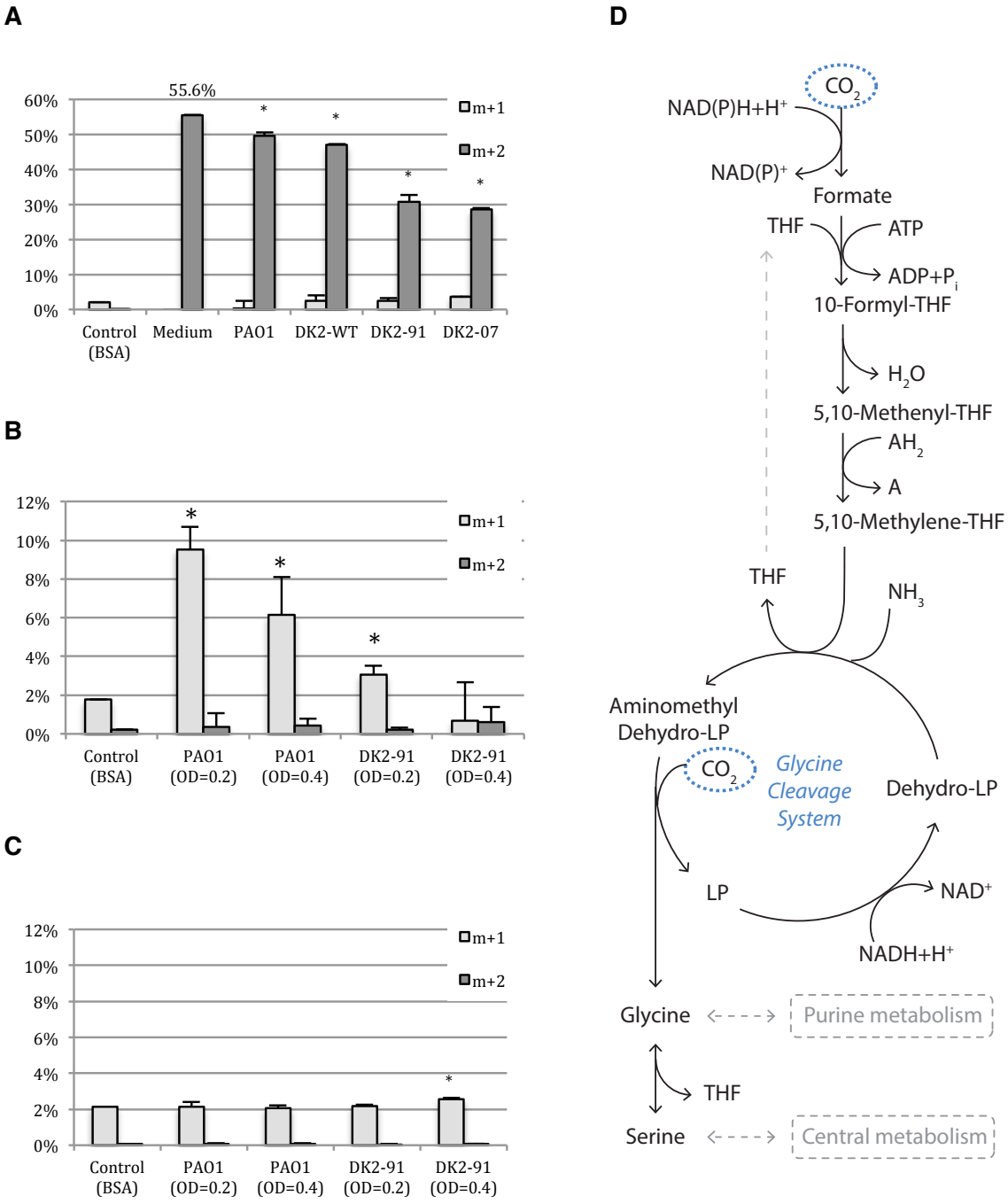
File contains each isolate-specific model in spreadsheet form with all applied gene and reaction constraints (and their corresponding reaction bounds as interpreted via gene-to-protein relationships).

Paper 2: Figures

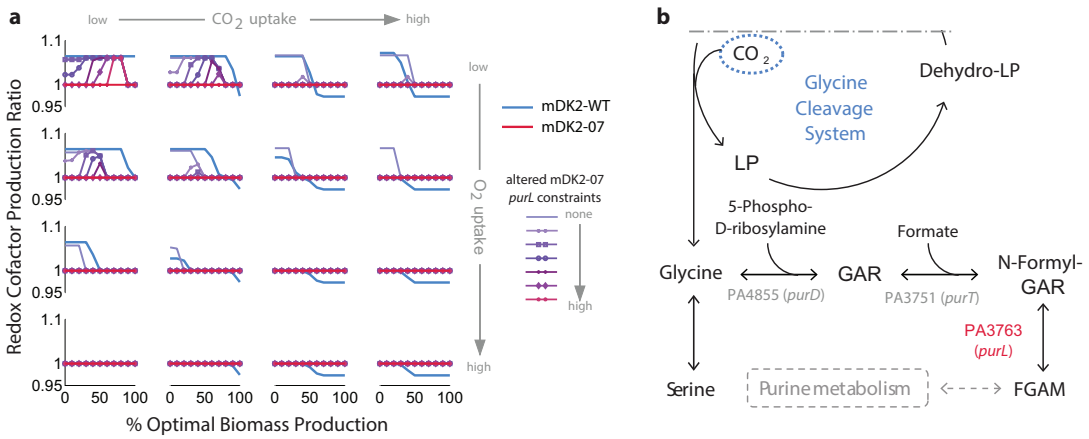
Paper 2 - Figure 1



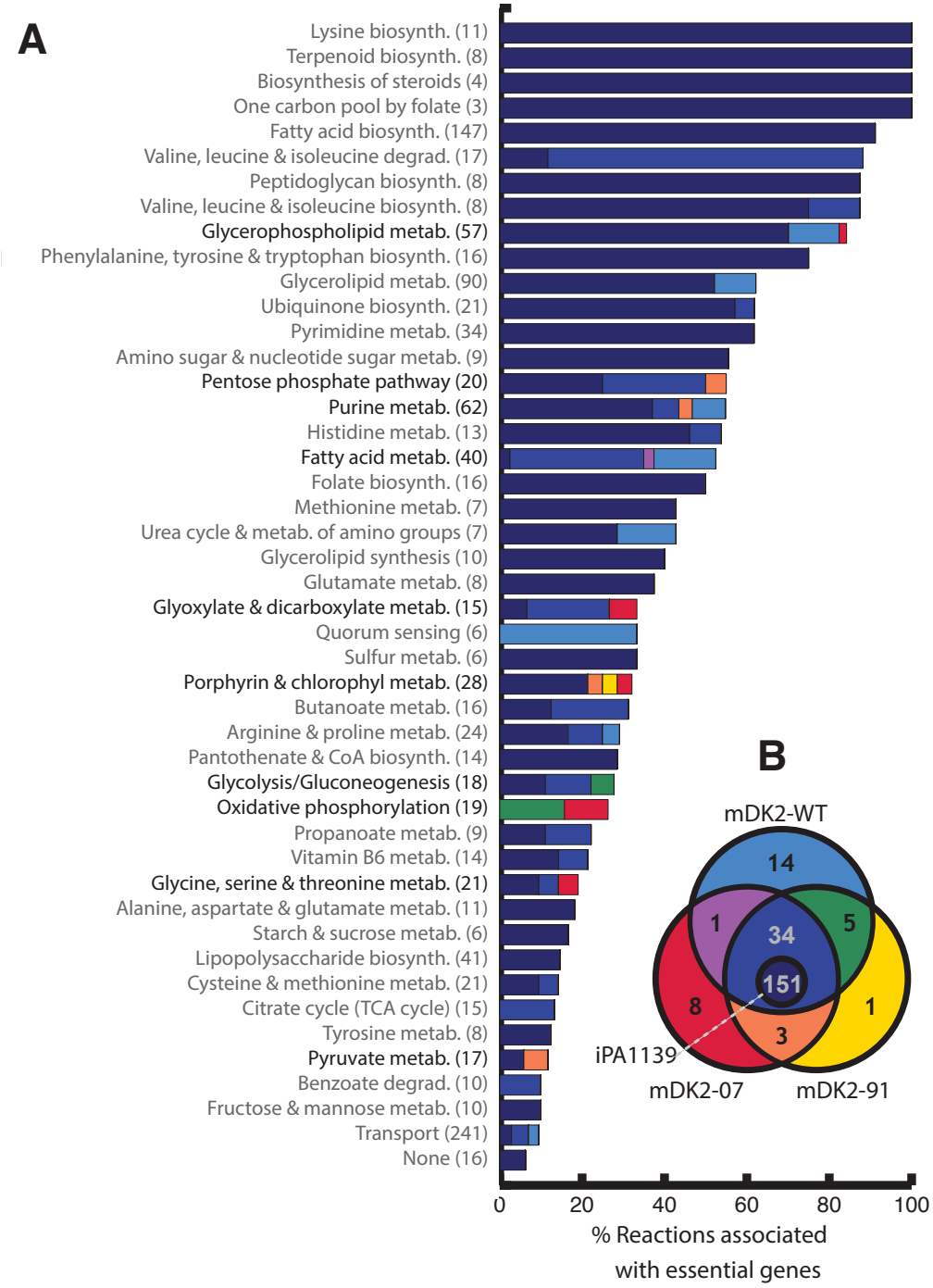
Paper 2: Figure 2



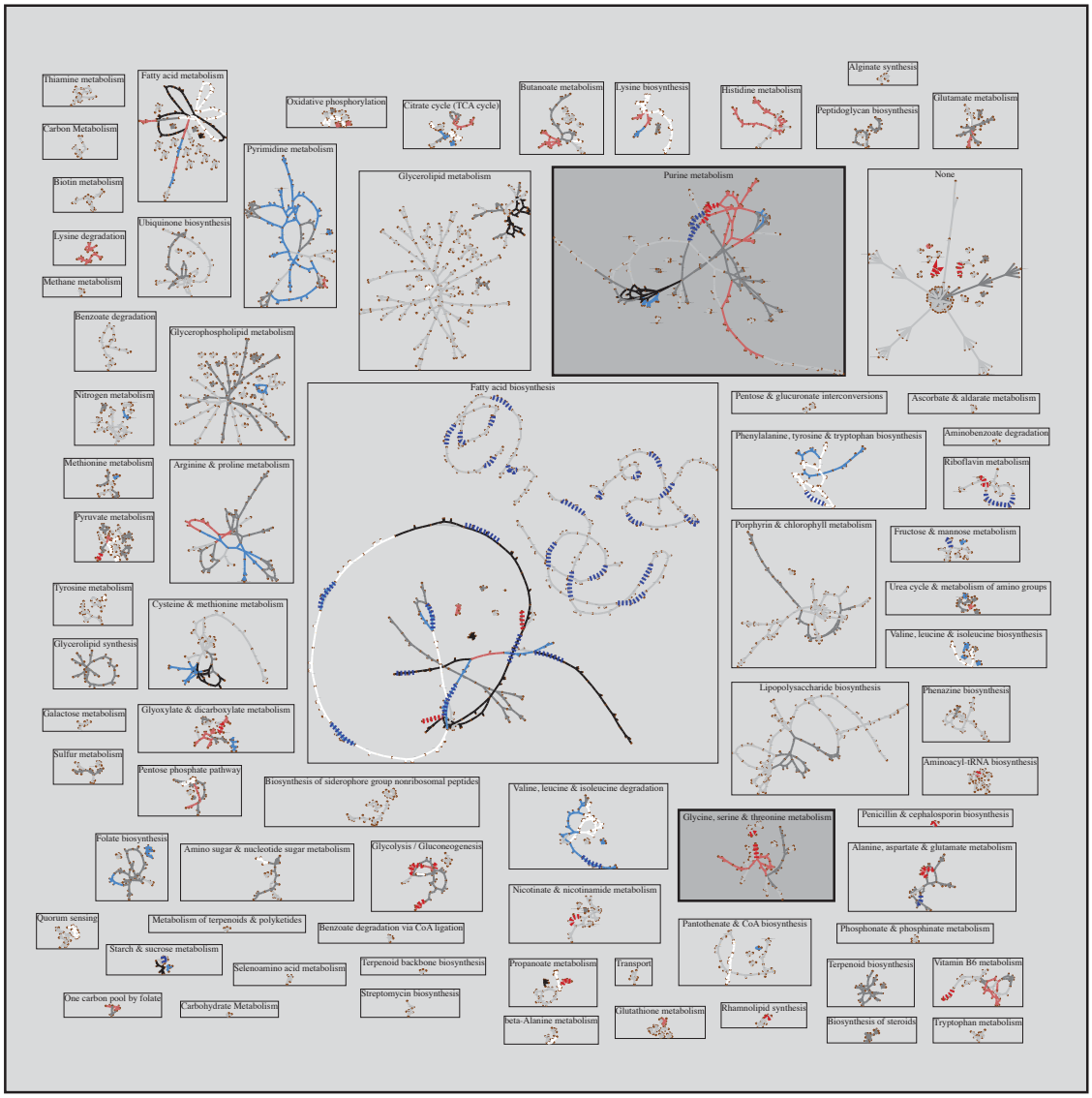
Paper 2 - Figure 3



Paper 2 - Figure 4



Paper 2 - Figure 5



Paper 2: Supplemental Text S1

Thoegersen *et al*: Systems-based analysis of metabolic evolution during pathogen adaptation to the human host, Supplemental text S1

Supplemental text S1: Expanded materials and methods

GC-MS analysis for extracellular metabolites

The experimental procedure of labelling determination was modified from Kind *et al.* 2010 (1). Supernatants were centrifuged at 15000 g and 100µL supernatant lyophilized in 2-ml silanized glass vials, and then derivatised by 20 µL O-methylhydroxylamine in pyridine for 2 hrs, before adding 180 µL of N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane (Thermo Fischer) heating in an oven at 40 °C for 30 min. The samples were analysed by GC-MS on a Thermo Electron DSQII GCMS systems using the same parameters as described in (1) for their Agilent GC-MS, and peaks matched in the Fiehn Lib (Agilent technologies) using the AMDIS 2.71 (http://chemdata.nist.gov/mass-spc/amdis/downloads/). Reference standards of glucose, gluconate, 5-ketogluconate and 2-ketogluconate were co-analysed in the sequence with real samples for verification.

Proteinogenic amino acid analysis from ^{13}C-labeled biomass

Hydrolysis:

The pellet was resuspended in 600 µL of 6 M hydrochloric acid and the volume was transferred to a 2 mL glass vial. The vial was capped with an aluminium cap (able to withstand high temperatures) and kept at 105°C for hydrolysis overnight. After overnight hydrolysis the content of the vial was transferred to an Eppendorf tube and centrifuged at 15000 rpm for two minutes. Supernatant was transferred to two clean glass vials (280 µL each). The vials were dried for three hours at 105°C without caps. After drying one of the vials was capped and stored at -80°C for

backup. The other vial was added 200 µL of milliQ water and vortexed for 30 seconds. Another 800 µL of milliQ water was added followed by vortexing. A control sample containing bovine serum albumin (BSA) was included to test if the hydrolysis step was completed successfully.

Purification:

The biomass hydrolysate was loaded on a cat-ion exchange solid phase extraction column packed in a 1–ml syringe (200 mg Dowex 50W X8, 200- 400 mesh, H^+-form, Sigma-Aldrich, St. Louis, MO), that had been conditioned by 1 ml methanol and 1 ml water, and the sample was passed through by gravity. Waste was discarded. The sample was washed with 1 mL of ethanol in water (1:1). 0.2 mL of 1 M NaOH was added to increase the pH of the column and waste was discarded. A 2 mL glass vial was placed under the column to collect the purified amino acids. 1 mL of a mixture of 1% (wt/v) NaOH in saline, ethanol and pyridine in a 9:5:1 proportion was added and the eluate was collected. The content was divided into two parts, 500 µL in an Eppendorf tube for ethylchloroformate (ECF) derivatization and 500 µL in a glass vial for N-dimethyl-amino-methylene-methyl-esters (DMFDMA) derivatization respectively. The samples were kept at -20°C until derivatization. A control sample containing a mixture of amino acids was included to test if the purification step was completed successfully.

ECF Derivatization:

50 µL of ethylchloroformate was added to the 500 µL SPE column eluate. Pipetting in and out using a 1 mL pipette followed by a gentle vortexing gently mixed the content. The Eppendorf tube was uncapped to release the pressure. This step was repeated until no CO_2 was observed. 5 additional µL of ECF was added followed by vortexing and release of pressure. 200 µL of propyl

acetate was added, the tube was vortexed for 30 seconds and pressure was released. 50 µL of 1 M HCl was added followed by vortexing and release of pressure. The fluid was allowed to separate for 1 minute. Thereafter 175 µL of the upper organic layer was transferred to a new Eppendorf tube. A small amount of anhydrous $NaSO_4$ or $MgSO_4$ was added followed by vortexing. The supernatant was transferred to a 2 mL glass vial and kept at -20°C until GC-MS analysis.

DMFDMA Derivatization:

200 µL 1 M HCl was added to the 500 µL SPE column eluate and mixed well. The vial was kept for drying for 2 to 4 hours at 105°C without cap. The vial was allowed to cool down for ten minutes. 200 µL DMFDMA and 200 µL acetonitrile was added to the vial. The vial was capped with a screw cap and kept for derivatization at 100°C for 20 minutes. After derivatization the vial was placed at -20°C for 10 min. The supernatant was transferred to an Eppendorf tube and centrifuged at 15.000 rpm for 2 min. The supernatant was transferred to a new glass vial with a screw cap and kept at -20°C until GC-MS.

GC-MS analysis of proteinogenic amino acids

Sample were analysed by GC-MS one an Agilent 6890 gas chromatograph (Agilent Technologies, Waldbronn, Germany) coupled to an Agilent 5973 quadruple MS run in electron impact ionization (EI^+) mode using an electron energy of 70 eV. The GC was equipped with a 4.0 mm i.d. Siltek gooseneck splitless deactivated liner (Restek, Bellefonte, PA, USA), and a Supelco (Bellefonte, PA, .US) Equity®-1701 (15 m, 0.25 mm i.d., 0.25 µm film) column. Helium was used as carrier gas at a constant linear gas velocity of 38 cm/s. Transfer line

temperature was 280ºC, quadruple temperature 150 ºC and MS source 230 ºC. The GC-MS system was controlled from Agilent MSD Chemstation v. D.01.02.16, and auto tuned for prior to every sequence. Samples of 1 µL was injected using a Combi PAL autosampler (CTC Analytics AG, Zwingen, Switzerland).

Analysis of amino acid-ECF derivatives was done at an injection temperature of 220ºC, and oven temperature was initially held at 75 °C for one min. Hereafter the temperature was raised 40 °C min^{-1} until 165 °C, then 4 °C min^{-1} until 190 °C and then 40 °C min^{-1} to 240 °C. At the end, temperature was increased to 260 °C at 4 °C min^{-1} and held constant for 4 minutes.

Analysis of the amino acid-DMFDMA derivatives was done at an injection temperature of 230ºC, and oven temperature was initially held at 60 °C for one min. Hereafter the temperature was raised at 20 °C min^{-1} until 130 °C, then 4 °C min^{-1} until 150 °C and 40 °C min^{-1} to 260 °C and held constant for 4.25 minutes.

Construction of isolate-specific genome-scale metabolic models

Raw expression levels from microarrays were used to develop proposed 'off' and 'on' gene activity levels using 25^{th} and 75^{th} percentile cutoffs of the expression data similar to methods described by Machado and Herrgard (2014) (2) These gene levels were converted into tri-valued logic levels ('off' – 0, 'unconstrained' – 1, and 'on' – 2) as the input for the TIGER implementation of iMAT. Different levels of SNP constraints were also used, ranging from minor impact (silent and SIFT-predicted tolerated missense SNPs), moderate impact (missense SNPs with SIFT-predicted functional impact), and maximum impact (nonsense SNPs). In order to integrate these datasets before iMAT was used to create strain-specific models, any Boolean gene-to-protein-to-reaction (GPR) relationship that incorporated a gene associated with a SNP

was manually evaluated in the context of the gene expression levels. If only the SNP-affected gene was associated with the reaction, the activity of the connected reaction was limited by modifying the reaction bounds. If the GPR was a more complex Boolean statement involving multiple genes (gene duplications, isozymes, or multiple subunits of an enzyme), the GPR was evaluated to see whether any genes were present that could compensate for the affected function of the SNP-associated gene. If these compensatory genes also had 'off' expression levels, the SNP-based constraint was applied. If the compensatory genes were unconstrained or 'on', the SNP-based constraint was not applied. Instead of reducing the potential activity of SNP-targeted reactions by their base model bounds (usually -1000 to 1000 for reversible reactions and 0 to 1000 for irreversible reactions), we conducted flux variability analysis of the base model at 100% biomass production to calculate the normal range of activity of each reaction in glucose minimal medium conditions. Any minor impact SNP being implemented resulted in a 10% reduction of the FVA-predicted base activity range enforced via reaction bounds while a moderate impact SNP resulted in a 50% reduction applied in the same manner. SNPs implemented with maximum impact resulted in associated reactions being turned entirely off via modification of reaction bounds. This GPR-based evaluation of SNP and expression levels is available in Dataset S2.

The above SNP integration method was applied to each strain-specific model prior to the use of the TIGER implementation of iMAT. Using an objective function threshold of 10% of the maximum and Gurobi 5.6.2 as the solver, iMAT predicted new sets of genes that could feasibly be turned 'off' or 'on' while maintaining production of biomass at 10% in each SNP-constrained isolate-specific model. The 'off' genes were inactivated in the model. The predicted 'on' genes

were implemented by applying a lower bound constraint of 0.001 or -0.001 to ensure a minimum level of activity in the appropriate direction of reaction activity. Reaction direction was evaluated via FVA, and if there was not a clear preference for direction of activity (for example, the FVA max and min indicated the reaction was fully reversible (a range of -0.001 to 0.001 or larger)), then the 'on' minimum constraint was not applied to avoid inappropriate/unsupported bias in reaction directionality. This evaluation meant that it was not feasible to apply all constraints predicted by iMAT, and a summary of the gene constraints and the difference between predicted and applied isolate-specific model constraints is presented in Figure S5.

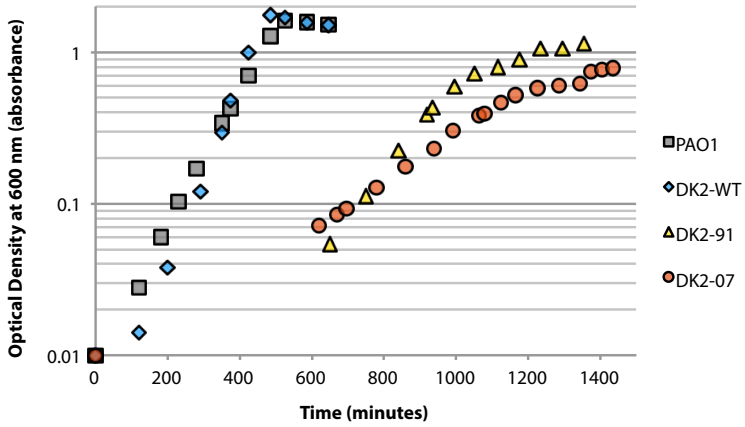
1. **Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, Fiehn O**. 2010. FiehnLib – mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gaschromatography/mass spectrometry. Anal. Chem. **81**:10038–10048.

2. **Machado D, Herrgård M**. 2014. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput. Biol. **10**:e1003580.

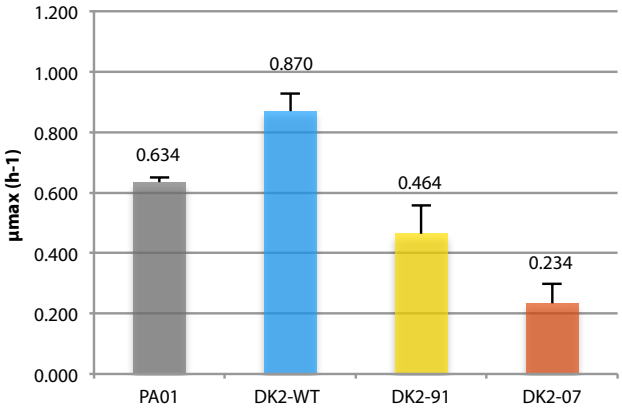
Paper 2: Supplemental Figures S1-S5

Paper 2 - Supplemental Figure S1

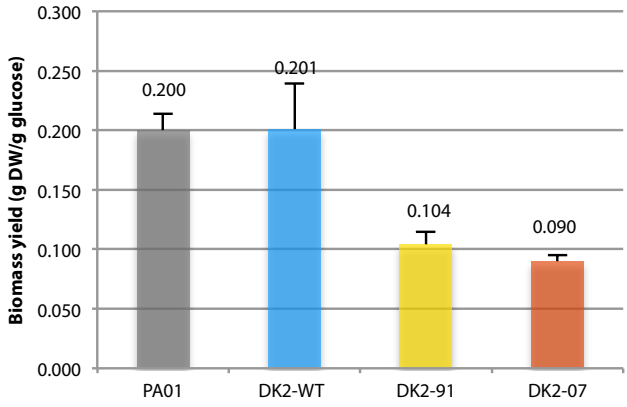
A



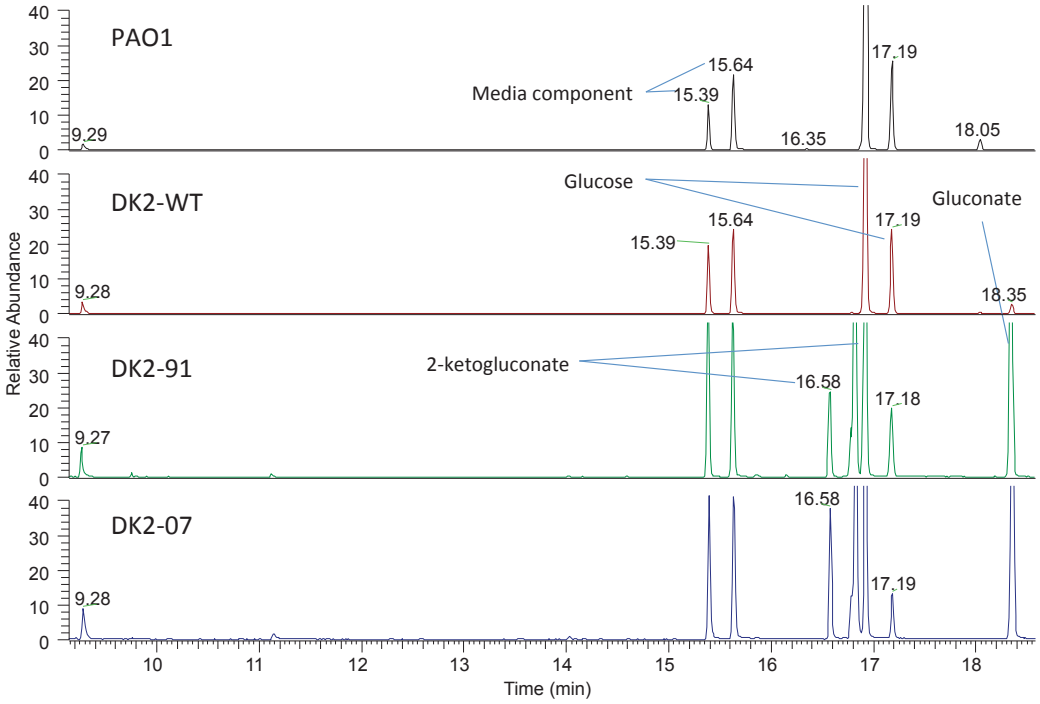
B



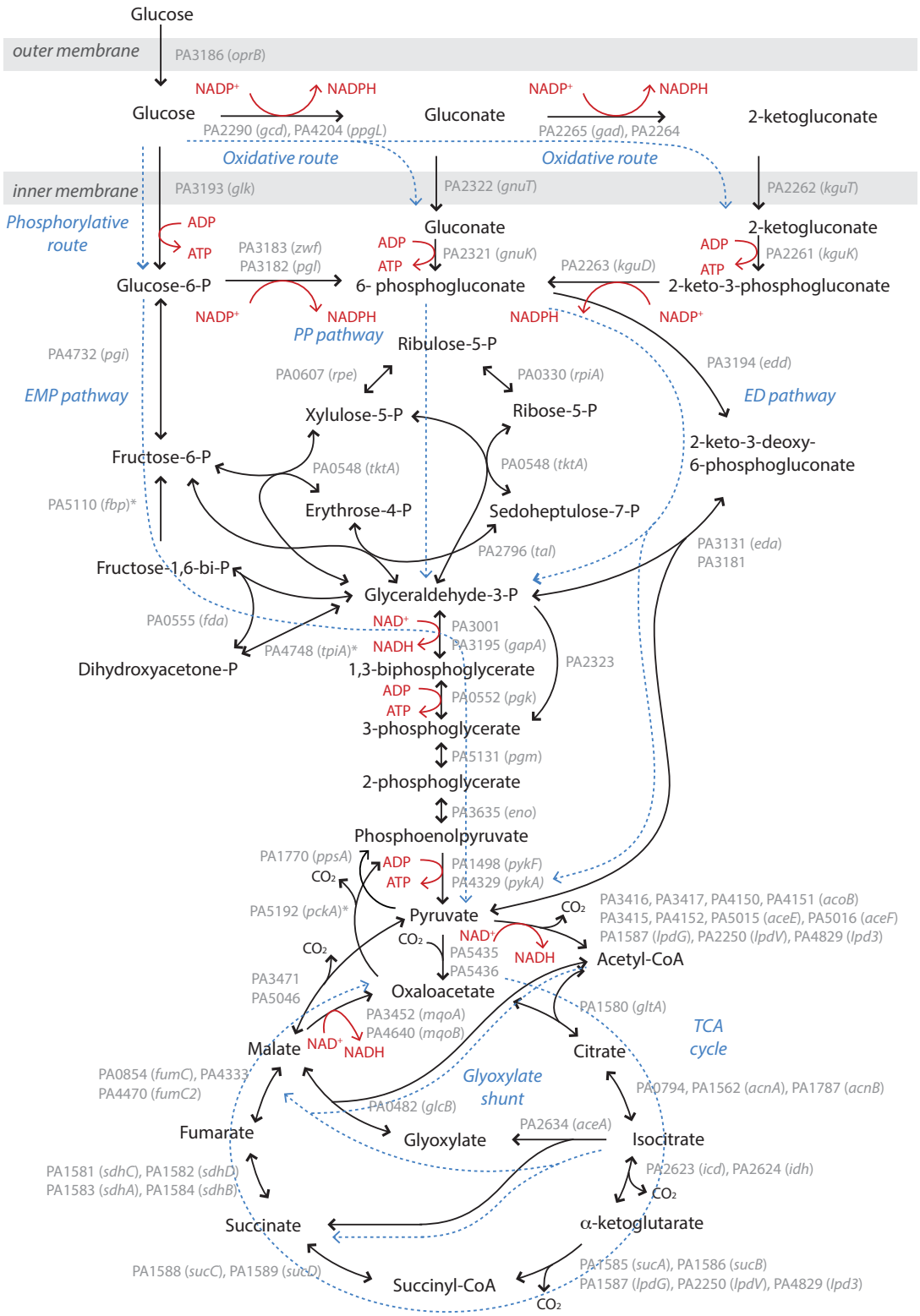
C



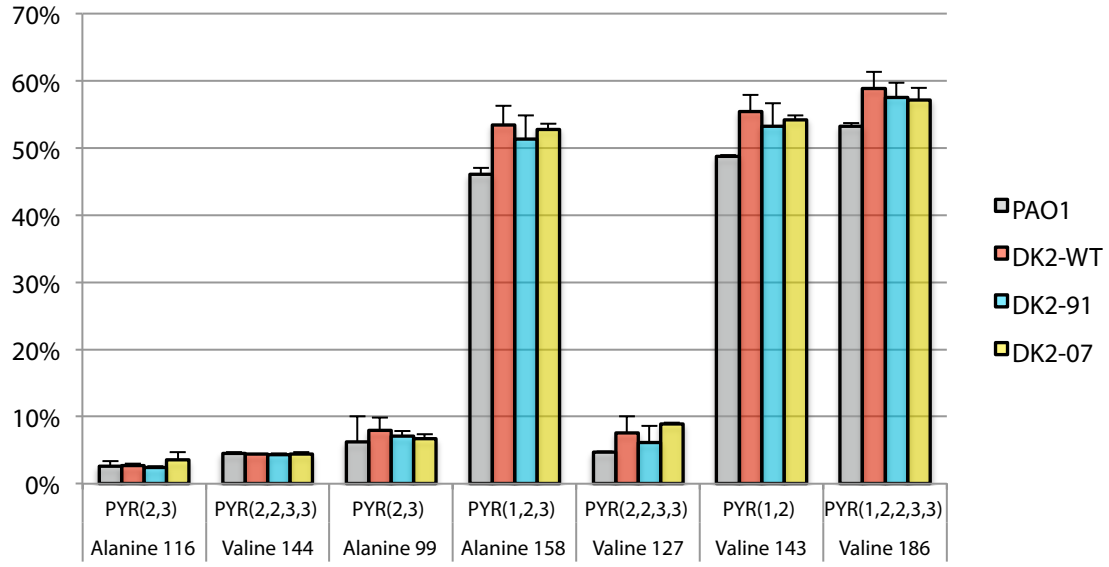
Paper 2 - Supplemental Figure S2



Paper 2 - Supplemental Figure S3



Paper 2 - Supplemental Figure S4



Paper 2 - Supplemental Figure S5

		DK2-WT	DK2-91	DK2-07
		constrained genes		
Gene Exp.	*off*	323	287	239
	on	324	271	260
SNPs	*off*	0	2	2
	50%	3	5	9
	90%	11	8	14
		TIGER-predicted constraints		
iMAT	*off*	299	273	221
	on	174	153	147
		constrained model reactions		
combined integration:		364	296	271

Paper 2: Supplemental Data Sets S1-S4

Available online at (until the paper is published):

https://www.dropbox.com/sh/645k1g7hji2eaan/AAAtkE6AWJ0soZoyGOB1SFLwa?dl=0

Chapter 7

Discussion

The main objective of this PhD project was to gain more knowledge of the adaptation process of human pathogens during chronic infections through a systems biology approach.

In **Paper 1** we use archetypal analysis to extract phenotypes of adapted *P. aeruginosa* isolates from global gene expression data sets derived from five diverse studies of *P. aeruginosa*. We are able to group *P. aeruginosa* isolates based on adaptation level and to represent the diverse data points as representative archetypes. The characterization of archetypes reveals that one archetype represents the mucoid phenotype and another archetype represents the hypermutator phenotype, both of which are frequently observed phenotypes isolated from chronic infections. Further characterization of the archetypes uncovers typical differential gene expression between isolates from early and chronic infections respectively.

In **Paper 2** we use long-term chronic infections of cystic fibrosis airways by *P. aeruginosa* as a model system for systematic analysis of how evolution shapes the metabolism of infecting bacteria. The study sheds light on specific metabolic pathways that have not previously been considered important for pathogen adaptation and persistence. We show that our approach of integrating high-throughput data into genome scale models can deliver novel insight into within-host evolution of bacterial pathogens.

The key contributions from **Paper 2** are:

- We provide direct evidence for metabolic activity towards fixation of carbon dioxide into glycine - a new discovered adaptive phenotype of the opportunistic pathogen *P. aeruginosa.*
- We present a novel methodological framework of integrating multiple data sets into genome scale metabolic models facilitating a systems level characterization of metabolism during adaptation.
- Through our combined experimental and computational approach, we can (i) connect the observed changes in metabolism to altered redox balance during adaptation (ii) predict which single nucleotide polymorphism is contributing to this change, (ii) develop specific lists

of predicted essential genes for the host-adapted isolates, which can serve as alternative therapeutic targets and (iv) prioritize impacted subsystems for future investigation.

The finding of the metabolic shift through the glycine cleavage system in **Paper 2** is a new discovery, which is made possible through our approach with labeling experiments and computational modeling. If we consider the genomic data or transcriptomic data alone, we do not find any SNPs or any differential gene expression in the genes coupled to the glycine cleavage system. We do identify a SNP in a neighboring system (purine metabolism), but connecting this SNP to the glycine cleavage system is unique to this study. We have identified a potential important pathway through the study of metabolism in *P. aeruginosa*, where no clear identification was possible through investigation of genomic or transcriptomic data alone. I think this is a great example of a successful outcome of data integration and metabolic modeling, where new discoveries appear from data sets that were not made by analyzing the data sets individually.

Is the glycine cleavage system also affected in other studies of P. aeruginosa adaptation?

The use of the *P. aeruginosa* DK2 lineage to represent general *P. aeruginosa* adaptation or even general pathogen adaptation may be argued. The results from **Paper 1** show that the DK2 lineage has adaptive traits common to other adapted lineages of *P. aeruginosa*. In addition to that, the DK2 lineage has undergone parallel evolution in multiple patients where it has outcompeted other *P. aeruginosa strains* (Jelsbak *et al*, 2007; Yang *et al*, 2011). I think that these examples are justifying the choice of DK2 to represent at least a common path of *P. aeruginosa* adaptation. Of course, it is always desired to confirm whether the specific adaptive changes found for the DK2 lineage (*e.g*. the altered glycine cleavage system) are observed for other lineages of *P. aeruginosa.* Interestingly, adaptive mutations have recently been discovered in a gene of the glycine cleavage system (*gcvP1*) among other clinical isolates of *P. aeruginosa* (Feliziani *et al*, 2014) indicating that this system is also altered in other *P. aeruginosa* lineages where it is even a direct target of genetic adaptation. The altered glycine cleavage system is therefore not unique to the DK2 lineage. Whether the identified adaptive mechanisms for within-host persistence are valid for other pathogens requires further investigation.

In the introduction I mentioned a comprehensive study of *P. aeruginosa* adaptation focusing on genomics (Marvig *et al*, 2015). With our gained attention to the glycine cleavage system and surrounding metabolic pathways, it could be interesting to evaluate whether an increased frequency of mutations within the defined set of metabolic pathways is present in the collection of 474 whole-genome sequences of *P. aeruginosa* clinical CF isolates from Marvig et al. The analysis is available in Appendix

A. In order to evaluate this we need to define which genes belong to the metabolic subsystem of interest. We chose to use Figure 5 as the delimiter of the subsystem and hence focus on the 19 genes that are involved in the reactions included in Figure 5. None of the 19 genes that appear in Figure 5 were identified as pathoadaptive in the study by Marvig et al (Marvig *et al*, 2015). However, if we count the number of mutations (including missense mutations, silent mutations and indels) that appear in the 19 genes we find that the system is significantly enriched for mutations compared to the average mutation frequency of all genes in the 36 lineages (Poisson distribution test, p =0.05). This means that the metabolic subsystem in Figure 5 is a target of genomic adaptation although not identified by the genomic study alone. I think this example is a step forward in moving from the genotype to the phenotype. Our modeling approach allows us to identify a subsystem of metabolic reactions that should be considered together and this now enable us to predict a new phenotype (altered glycine cleavage system) from the genomic data.

I wish to emphasize that the identification of the metabolic shift through the glycine cleavage system in itself is a major result of this PhD project. The finding of course gives rise to new project proposals, where the direct role of the glycine cleavage system connected to *P. aeruginosa* persistence could be investigated.

Is the metabolic shift through the glycine cleavage system relevant for other organisms?

Our increasing interest for the glycine cleavage system also inspired us to look for other disease cases, where the glycine cleavage system has shown to be important. As also mentioned in **Paper 2**, the enzyme glycine decarboxylase (part of the glycine cleavage system) was previously identified as a promoter of cellular transformation in cancer cells and the activity of this enzyme was correlated with poorer survival of patients with lung cancer (Zhang *et al*, 2012). A very interesting link between our study and these previously published observations is that in both cases (i) the lung environment has a high partial pressure

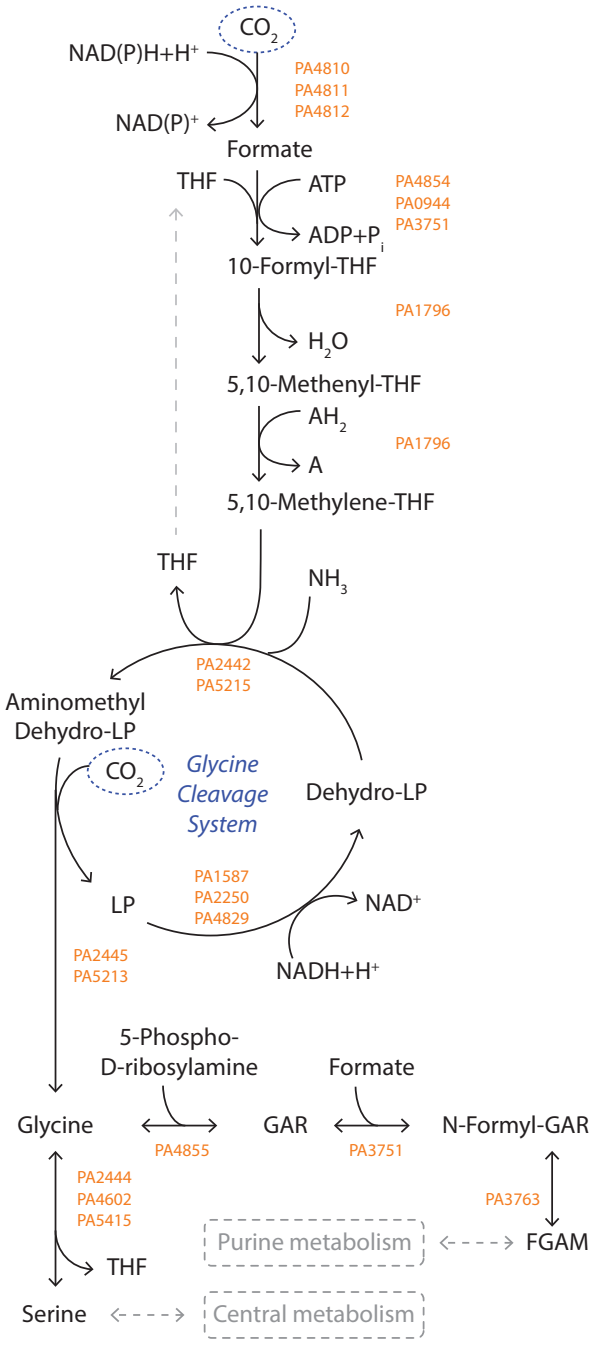


Figure 5. The glycine cleavage system in reverse. Two molecules of carbon dioxide are fixated into glycine - one of them via formate formation. Genes are assigned to each reaction. Figure adapted from (Bar-Even *et al*, 2012). Abbreviations: Reduced electron donor (AH_2), oxidized electron donor (A), tetrahydrofolate (THF), lipoyl protein (LP), glycinamide ribonucleotide (GAR), 5'-phosphoribosyl-formylglycinamidine (FGAM).

of CO_2, (ii) changes in metabolism indicate decreased oxidative phosphorylation, and (iii) changes are observed in glycine metabolism and the glycine cleavage system. A key difference is that we propose that the mechanism of the glycine cleavage system is to regenerate reducing equivalents through glycine synthesis, whereas in the study by Zhang *et al.* (Zhang *et al*, 2012), the glycine cleavage system is considered irreversible in the direction of glycine cleavage. However, the system is bi-directional (Bar-Even *et al*, 2012) and we suggest that the function of the glycine cleavage system in cancer cells could be similar to the mechanisms found in *P. aeruginosa* where it is operating in the direction of glycine synthesis. Carbon fixation through the glycine cleavage system in cancer cells

could be a way to fuel glycolysis with redox compounds thereby achieving the Warburg effect, a well-known phenomenon in cancer cells characterized by high glycolytic activity and low (or normal) level of oxidative phosphorylation (Chen *et al*, 2015). Future studies could examine the potential reversal of the glycine cleavage system by evaluating growth of malignant cancer cells provided with labeled carbon dioxide or bicarbonate to evaluate whether a similar redox-related driving mechanism is associated with lung cancer metabolism.

Can we derive what the driving force of selection is in the CF lung environment?

The often observed slow-growth phenotype among the *P. aeruginosa* isolates from chronic patients indicate that optimization of biomass yield may not be the strongest selective pressure in the lung environment, although it is contrary to the general assumption that bacteria have been optimized evolutionarily for growth (Oberhardt *et al*, 2008; Schuetz *et al*, 2007). As described by Shoval et al (2012), a bacterial phenotype cannot be optimal, at the same time, for two different tasks such as rapid growth and survival, but the bacteria will meet a tradeoff between the two objectives (Shoval *et al*, 2012). The question is if we can identify the important tasks that contribute to the fitness of the bacteria in the CF lung environment.

When the first genome-scale metabolic model of *P. aeruginosa* was published in 2008, it was stated that some future application of that model could be to model different hypothesis about selective pressures in the lung and to analyze the causes of these selective pressures (Oberhardt *et al*, 2008). I think that our work presented in **Paper 2** is an example of this. In **Paper 2** we simulate an effect of carbon dioxide and oxygen levels on redox potential between two isolates where we find that for the adapted isolate, the redox potential stays balanced despite fluctuating carbon dioxide and oxygen levels, which could mean decreased sensitivity to surrounding conditions including oxidative stress. The work therefore has led to the hypothesis:

Hypothesis:

Pseudomonas aeruginosa is able to resist oxidative stress in the cystic fibrosis lung environment due to metabolic reprogramming of the glycine cleavage system.

It is possible that balancing redox potential in the cells is one of the driving forces of selection in the CF lung environment. Future work could include measurements of redox potential under various stress conditions (including oxidative stress) and comparisons between a wild type *P. aeruginosa* and an isogenic *purL* mutant (the mutation that we predict is causing the metabolic shift). In **Paper 2,** we

also show that the *purL* mutation causes an *in silico* growth defect and therefore the potential advantage of resisting oxidative stress through redox balancing seems to pay a price on growth.

Can we extract in vivo phenotypes from in vitro data?

One often argued question is how we can relate laboratory experiments to reality. One advantage with our model system is that the evolutionary process of *P. aeruginosa* that we investigate has taken place *in vivo* that means in the natural host environment within the CF lung. The bacterial isolates collected from the CF patients should represent a snapshot of the stage of adaptation at the time of isolation. However, when we characterize the bacterial isolates, we make a transition from *in vivo* (the host) to *in vitro* (the laboratory). When we evaluate *in vitro* phenotypes between bacterial strains we need to consider that these phenotypes might be affected by the laboratory setup and one challenge is how we mimic the authentic environmental conditions in the laboratory. The nutrient composition of CF sputum has been characterized in order to make a synthetic medium that should represent CF sputum (Palmer *et al*, 2007) but many parameters of the CF lung ecology remains unknown. Also, these conditions are not always applicable to the experimental investigations. One example is isotope-labeling experiments where you want to avoid presence of unlabeled carbon sources and for our study in **Paper 2** this is accomplished by growing cells in glucose minimal medium, which is different from the nutrient rich sputum medium. Another aspect is the mode of growth and whether the bacteria grow as static cells, planktonic cells or maybe as biofilm structures, which can also affect other *in vitro* measured phenotypes.

In our experiment with labeled glucose we aim at investigating central metabolism. Central metabolism is definitely affected by which nutrients are present and measured activities of pathways through central metabolism are most likely different from the actual metabolism *in vivo*. However, our focus is on the metabolic differences between bacterial isolates representing different stages of adaptation. Our main interest is not the actual reaction activities, which are subject to the bias by the dissimilar *in vitro* environment, but rather the pathways that have differential activities between bacterial isolates. The results connected to these differential activities should more likely be due to genetic differences in the bacteria rather than regulatory effects of the surrounding environment. We thereby attempt to cancel out the *in vitro* bias through our comparative analysis.

In **Paper 1** we also deal with the issue of laboratory conditions. Again, our focus is on adaptation and we merge data from five distinct studies with varied laboratory conditions applied. We identify two distinctive phenotypes (the mucoid phenotype and the hypermutator phenotype) in our transcriptomic data analysis across experimental conditions. The conditions in which the bacteria are

cultivated will affect the transcriptome, but we show that through appropriate choice of analytical method, we can extract the relevant phenotypes from the data set and neglect the bias from the experimental setup. The results from **Paper 1** also show that other methods including PCA fail to identify these patterns. This can be caused by the fact that PCA is an algorithm that directs our attention to most variance in data and this method can therefore be very sensitive to variations caused by diverse experimental conditions. It is therefore important to take these issues into consideration in design of experiments and connected analytical frameworks.

Our finding of the shared gene expression pattern among hypermutators in **Paper** 1 suggests that the hypermutators have similar adaptive phenotypes. One interesting observation from **Paper 1** is that the transcriptome profiles of the hypermutators compared to their non-hypermutator relatives are less different from the average transcription profile for all samples included in the study. I suggest that this could be caused by decreased sensitivity to the change in environmental conditions. Mutations in global regulators have previously been associated with adaptation of *P. aeruginosa* (Yang *et al*, 2011; Damkiær *et al*, 2013; Smith *et al*, 2006) and maybe one selective advantage is to diminish the regulatory response to environmental changes.

Concluding discussion and future perspectives

We have successfully applied system biology approaches to the study of *P. aeruginosa* adaptation. We have developed isolate-specific genome-scale metabolic models that were able to identify particular metabolic subsystems that are subject to changes during adaptation of *P. aeruginosa* in the CF lung environment. In addition to that the analysis has identified genes that most likely become essential during adaption. Both results can be valuable as targets for future intervention. Multiple studies have reported parallel evolution of different *P. aeruginosa* lineages inside the CF lung (Huse *et al*, 2010; Weigand & Sundin, 2012; Yang *et al*, 2011; Marvig *et al*, 2013) and therefore I think that it is most likely that the observed metabolic changes and essential genes are valid for other *P. aeruginosa* strains. Our focus on the glycine cleavage system revealed that this particular metabolic subsystem was also a target of adaptive mutations in other lineages of *P. aeruginosa.*

For future studies it could be interesting to model a core and pan metabolic capacity for multiple diverse *P. aeruginosa* strains similar to the study of core and pan metabolism of multiple *E. coli* strains by (Monk *et al*, 2013). The core metabolism would account for metabolic reactions that are present in all isolates, whereas the pan metabolism would be the overall metabolic capacity of *P. aeruginosa* accounting for all metabolic genes present in any *P. aeruginosa* strain. From such a study, it is possible that activities in some parts of metabolism can be related to successful estab-

lishment of chronic *P. aeruginosa* infections in CF patients. Another possible expansion on the modeling work could be to build multi-species metabolic models, where the interactions between for example *S. aureus* and *P. aeruginosa* are investigated.

One issue that remains a challenge to our use of systems biology is the trust in the metabolic models. When do we think the models are perfect enough to trust the predictions without experimental validation? Or will we ever reach a state where we can circumvent the need for experimental validation? And are our synthetic laboratory setups even more precise than our modeling results? Sometimes it can be difficult or technically infeasible to set up adequate validation experiments for model predictions. I think these model predictions are very interesting because they allow us to get insight into some parts of metabolism that we have so far not been able to characterize in the laboratory.

The genome-scale metabolic models will represent a simplification of cellular function, since the combination of metabolism, regulation and signaling that are network components of a living cell is much more complicated than we can model (Oberhardt *et al*, 2009). I think we should remember that creating perfect models of metabolism is not the ultimate aim. As described by Heinemann and Sauer (2010), the aim in systems biology is a global system understanding where the core interest is the general principles underlying a particular system rather than exact molecular mechanisms (Heinemann & Sauer, 2010). The aim is therefore to be able to provide new biological understanding through the applications of these models and I think that both correct and incorrect predictions can contribute to that. Incorrect predictions will assist in identifying parts of metabolism that we need to investigate further. Correct predictions of unknown cellular function will most likely help us reaching a knowledge level faster than if we only employed an experimental approach. I don't think we will or should pursue to reach a stage where experimental validation becomes negligible for genome-wide analyses. However, I do think that we will be able to validate subsets of metabolism so that we are able to trust future predictions within the same subsystem without experimental confirmation.

I started out this thesis with a quote: *"All models are wrong. Some are useful"* (Box, 1979). In an essay from 1993, Daniel Hillis elaborates on this: *"Those [models] that are most useful will probably not predict any particular experimental data, but instead they will give some surprising ideas about how something might work"* (Hillis, 1993). I think this is exactly what we should remember when we deal with modeling in systems biology.

Chapter 8

References

Aebersold R, Hood LE & Watts JD (2000) Equipping scientists for the new biology. *Nat. Biotechnol.* **18:** 359

Ando H, Miyoshi-Akiyama T, Watanabe S & Kirikae T (2014) A silent mutation in mabA confers isoniazid resistance on Mycobacterium tuberculosis. *Mol. Microbiol.* **91:** 538–547

Bar-Even A, Noor E & Milo R (2012) A survey of carbon fixation pathways through a quantitative lens. *J. Exp. Bot.* **63:** 2325–2342

Bartell JA, Yen P, Varga JJ, Goldberg JB & Papin JA (2014) Comparative metabolic systems analysis of pathogenic burkholderia. *J. Bacteriol.* **196:** 210–26

Becker SA & Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* **4:** e1000082

Bordbar A, Monk JM, King ZA & Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. **15:** 107–120

Box GEP (1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics*, Launer R & Wilkinson G (eds) pp 201–236. New York: Academic Press

Buchanan PJ, Ernst RK, Elborn JS & Schock B (2009) Role of CFTR , Pseudomonas aeruginosa and Toll-like receptors in cystic fibrosis lung inflammation. **:** 863–867

Burns JL, Emerson J, Stapp JR, Yim DL, Krzewinski J, Louden L, Ramsey BW & Clausen CR (1998) Microbiology of Sputum from Patients at Cystic Fibrosis Centers in the United States. **27:** 158–163

Del Castillo T, Ramos JL, Rodríguez-Herva JJ, Fuhrer T, Sauer U & Duque E (2007) Convergent peripheral pathways catalyze initial glucose catabolism in Pseudomonas putida: genomic and flux analysis. *J. Bacteriol.* **189:** 5142–52

Chen X, Qian Y & Wu S (2015) The Warburg effect: Evolving interpretations of an established concept. *Free Radic. Biol. Med.* **79:** 253–263

Ciofu O, Lee B, Johannesson M, Hermansen NO, Meyer P & Høiby N (2008) Investigation of the algT operon sequence in mucoid and non-mucoid Pseudomonas aeruginosa isolates from 115 Scandinavian patients with cystic fibrosis and in 88 in vitro non-mucoid revertants. *Microbiology* **154:** 103–13

Clatworthy AE, Pierson E & Hung DT (2007) Targeting virulence: a new paradigm for antimicrobial therapy. *Nat. Chem. Biol.* **3:** 541–8

Cooper TF, Rozen DE & Lenski RE (2003) Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. *Proc. Natl. Acad. Sci. U. S. A.* **100:** 1072–1077

Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E & Palsson BO (2001) Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci.* **26:** 179–186

Cutler A & Breiman L (1994) Archetypal Analysis. *Technometrics* **36:** 338–347

Cystic fibrosis foundation patient registry 2009 annual data report (2010) Bethesda, MD, USA

Damkiær S, Yang L, Molin S & Jelsbak L (2013) Evolutionary remodeling of global regulatory networks during long-term bacterial adaptation to human hosts. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 7766–71

Dettman JR, Rodrigue N, Aaron SD & Kassen R (2013) Evolutionary genomics of epidemic and nonepidemic strains of Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 21065–70

Döring G, Conway S, Heijerman HGM, Hodson ME, Høiby N, Smyth A & Touw DJ (2000) Antibiotic therapy against Pseudomonas aeruginosa in cystic fibrosis : a European consensus. *Eur Respir J* **16:** 749–767

Döring G, Flume P, Heijerman H & Elborn JS (2012) Treatment of lung infection in patients with cystic fibrosis: current and future strategies. *J. Cyst. Fibros.* **11:** 461–79

Dwyer DJ, Belenky PA, Yang JH, MacDonald IC, Martell JD, Takahashi N, Chan CTY, Lobritz MA, Braff D, Schwarz EG, Ye JD, Pati M, Vercruysse M, Ralifo PS, Allison KR, Khalil AS, Ting AY, Walker GC

& Collins JJ (2014) Antibiotics induce redox-related physiological alterations as part of their lethality. *Proc. Natl. Acad. Sci. U. S. A.* **111:** E2100–9

Edwards JS & Covert M (2002) Minireview Metabolic modelling of microbes : the flux-balance approach. **4:** 133–140

Edwards JS & Palsson BO (1999) Systems properties of the *Haemophilus influenzae Rd* Metabolic Genotype. *Cell Biol. Metab.* **274:** 17410–17416

Eisenreich W, Dandekar T, Heesemann J & Goebel W (2010) Carbon metabolism of intracellular bacterial pathogens and possible links to virulence. *Nat. Rev. Microbiol.* **8:** 401–12

Elena SF & Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4:** 457–469

Feist AM, Herrgård MJ, Thiele I, Reed JL & Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7:** 129–43

Feliziani S, Marvig RL, Luján AM, Moyano AJ, Di Rienzo J a, Krogh Johansen H, Molin S & Smania AM (2014) Coexistence and Within-Host Evolution of Diversified Lineages of Hypermutable Pseudomonas aeruginosa in Long-term Cystic Fibrosis Infections. *PLoS Genet.* **10:** e1004651

Folkesson A, Jelsbak L, Yang L, Johansen HK, Ciofu O, Høiby N & Molin S (2012) Adaptation of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective. *Nat. Rev. Microbiol.* **10:** 841–51

Foweraker J (2009) Recent advances in the microbiology of respiratory tract infection in cystic fibrosis. **:** 93–110

Friedman J, Tibshirani R & Hastie T (2009) The Elements of Statistical Learning - Data Mining, Inference, and Prediction 2nd ed. New York: Springer-Verlag

Gilligan PH (1991) Microbiology of Airway Disease in Patients with Cystic Fibrosist. **4:** 35–51

Govan JR & Deretic V (1996) Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia. *Microbiol. Rev.* **60:** 539–74

Gunawardena J (2014) Models in biology: 'accurate descriptions of our pathetic thinking'. *BMC Biol.* **12:** 29

Haggart CR, Bartell JA, Saucerman JJ & Papin JA (2011) Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods Syst. Biol.* **500:** 411–33

Harrison F (2007) Mini-Review Microbial ecology of the cystic fibrosis lung. *Microbiology* **153:** 917–923

Haufe C (2013) Why do funding agencies favor hypothesis testing? *Stud. Hist. Philos. Sci.* **44:** 363–374

Hauser AR, Jain M, Bar-Meir M & McColley SA (2011) Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev.* **24:** 29–70

Heinemann M & Sauer U (2010) Systems biology of microbial metabolism. *Curr. Opin. Microbiol.* **13:** 337–43

Hillis WD (1993) Why physicists like models and why biologists should. *Curr. Biol.* **3:** 79–81

Hindré T, Knibbe C, Beslon G & Schneider D (2012) New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nat. Rev. Microbiol.* **10:** 352–65

Hogardt M, Hoboth C, Schmoldt S, Henke C, Bader L & Heesemann J (2007) Stage-specific adaptation of hypermutable Pseudomonas aeruginosa isolates during chronic pulmonary infection in patients with cystic fibrosis. *J. Infect. Dis.* **195:** 70–80

Høiby N (2006) P. aeruginosa in Cystic Fibrosis Patients Resists Host Defenses, Antibiotics. *Microbe* **1:** 571–577

Høiby N, Frederiksen B & Pressler T (2005) Eradication of early Pseudomonas aeruginosa infection. **4:** 49–54

Hood L (2003) Systems biology : integrating technology , biology , and computation. *Mech. Ageing Dev.* **124:** 9–16

Hull J, Vervaart P, Grimwood K & Phelan P (1997) Pulmonary oxidative stress response in young children with cystic fibrosis. *Thorax* **52:** 557–560

Huse H, Kwon T, Zlosnik J & Speert D (2010) Parallel evolution in Pseudomonas aeruginosa over 39,000 generations in vivo. *MBio* **1:** e00199–10

Ideker T, Galitski T & Hood L (2001) A new approach to decoding life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2:** 343–372

Jelsbak L, Johansen HK, Frost A-L, Thøgersen R, Thomsen LE, Ciofu O, Yang L, Haagensen JAJ, Høiby N & Molin S (2007) Molecular epidemiology and dynamics of Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients. *Infect. Immun.* **75:** 2214–24

Jensen PA, Lutz KA & Papin JA (2011) TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC Syst. Biol.* **5:** 147

Jensen PA & Papin JA (2014) MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. *Bioinformatics* **30:** 1327–1328

Kjeldsen KR & Nielsen J (2009) In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network. *Biotechnol. Bioeng.* **102:** 583–97

Knowles MR & Boucher RC (2002) Innate defenses in the lung Mucus clearance as a primary innate defense mechanism for mammalian airways. **109:** 571–577

Kohanski MA, Dwyer DJ & Collins JJ (2010) How antibiotics kill bacteria: from targets to networks. *Nat. Rev. Microbiol.* **8:** 423–35

Kumar P, Henikoff S & Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4:** 1073–81

Lobel L, Sigal N, Borovok I, Ruppin E & Herskovits A a. (2012) Integrative Genomic Analysis Identifies Isoleucine and CodY as Regulators of Listeria monocytogenes Virulence. *PLoS Genet.* **8:**

De Lorenzo V (2014) From the selfish gene to selfish metabolism: Revisiting the central dogma. *BioEssays* **36:** 226–35

De Lorenzo V (2015) Pseudomonas aeruginosa: the making of a pathogen. *Environ. Microbiol.* **17:** 1–3

Lyczak JB, Cannon CL & Pier GB (2002) Lung Infections Associated with Cystic Fibrosis. **15:** 194–222

Machado D & Herrgård M (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Comput. Biol.* **10:** e1003580

Madigan MT & Martinko JM (2006a) Microbial interactions with humans. In *Brock Biology of Microorganisms* pp 701–726. London: Pearson Prentice Hall

Madigan MT & Martinko JM (2006b) Microorganisms and microbiology. In *Brock Biology of Microorganisms* pp 1–20. London: Pearson Prentice Hall

Marvig RL, Johansen HK, Molin S & Jelsbak L (2013) Genome analysis of a transmissible lineage of pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genet.* **9:** e1003741

Marvig RL, Sommer LM, Molin S & Johansen HK (2015) Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nat. Genet.* **47:** 57–65

Mathee K, Ciofu O, Sternberg C, Lindum PW, Campell JIA, Jensen P, Johnson AH, Givskov M, Ohman DE, Molin S, Høiby N & Kharazmi A (1999) Mucoid conversion of Pseudomonas aeruginosa by hydrogen peroxide: a mechanism for virulence activation in the cystic fibrosis lung. *Microbiology* **145:** 1349–1357

McDermott JE, Yoon H, Nakayasu ES, Metz TO, Hyduke DR, Kidwai AS, Palsson BO, Adkins JN & Heffron F (2011) Technologies and approaches to elucidate and model the virulence program of salmonella. *Front. Microbiol.* **2:** 121

Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N & Orth JD (2013) Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci.* **110:** 20338–20343

Mørup M & Hansen LK (2012) Archetypal analysis for machine learning and data mining. *Neurocomputing* **80:** 54–63

Nanchen A, Fuhrer T & Sauer U (2007) Determination of Metabolic Flux Ratios From 13 C-Experiments and Gas Chromatography–Mass Spectrometry Data. *Methods Mol. Biol.* **358:** 177–197

Oberhardt MA, Goldberg JB, Hogardt M & Papin JA (2010) Metabolic network analysis of Pseudomonas aeruginosa during chronic cystic fibrosis lung infection. *J. Bacteriol.* **192:** 5534–48

Oberhardt MA, Palsson BØ & Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5:** 320

Oberhardt MA, Puchałka J, Fryer KE, Martins dos Santos VAP & Papin JA (2008) Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. *J. Bacteriol.* **190:** 2790–803

Ohman DE & Chakrabarty AM (1982) Utilization of Human Respiratory Secretions by Mucoid Pseudomonas aeruginosa of Cystic Fibrosis Origin. *Infect. Immun.* **37:** 662–669

Oliver A & Mena A (2010) Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clin. Microbiol. Infect.* **16:** 798–808

Oliver AM & Weir DM (1985) The effect of Pseudomonas alginate on rat alveolar macrophage phagocytosis and bacterial opsonization. *Clin. exp. Immunol.* **59:** 190–196

Orth JD, Thiele I & Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* **28:** 245–248

Overbeek R, Begley T, Butler RM, Choudhuri J V., Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33:** 5691–5702

Palmer KL, Aye LM, Whiteley M, Al PET & Acteriol JB (2007) Nutritional Cues Control Pseudomonas aeruginosa Multicellular Behavior in Cystic Fibrosis Sputum †. **189:** 8079–8087

Palmer KL, Mashburn LM, Singh PK, Whiteley M & Al PET (2005) Cystic Fibrosis Sputum Supports Growth and Cues Key Aspects of Pseudomonas aeruginosa Physiology. *J. Bacteriol.* **187:** 5267–5277

Puchałka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, Timmis KN, Papin JA & Martins dos Santos VAP (2008) Genome-scale reconstruction and analysis of the Pseudomonas putida

KT2440 metabolic network facilitates applications in biotechnology. *PLoS Comput. Biol.* **4:** e1000210

Rahme LG, Stevens EJ, Wolfort SF, Shao J, Ronald G, Ausubel FM, Tompkins RG & Ausubel FM (1995) Common virulence factors for bacterial pathogenicity in plants and animals. *Science* **268:** 1899–902

Ramos J-L (2004) Pseudomonas - Genomics Life Style and Molecular Architecture New York: Kluwer Acacemic/Plenum

Rau MH, Hansen SK, Johansen HK, Thomsen LE, Workman CT, Nielsen KF, Jelsbak L, Høiby N, Yang L & Molin S (2010) Early adaptive developments of Pseudomonas aeruginosa after the transition from life in the environment to persistent colonization in the airways of human cystic fibrosis hosts. *Environ. Microbiol.* **12:** 1643–58

Roweis S & Ghahramani Z (1999) A unifying review of linear gaussian models. *Neural Comput.* **11:** 305–345

Schuetz R, Kuepfer L & Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol. Syst. Biol.* **3:** 119

Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K & Alon U (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science (80-.).* **1157:** 1157–60

Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, Miller SI, Ramsey BW, Speert DP, Moskowitz SM, Burns JL, Kaul R & Olson M V (2006) Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proc. Natl. Acad. Sci. U. S. A.* **103:** 8487–92

Sundin GW & Weigand MR (2007) The microbiology of mutability. *FEMS Microbiol. Lett.* **277:** 11–20

Thøgersen JC (2010) Metabolic Network Analysis of Glucose Metabolism for Clinical Isolates of Pseudomonas aeruginosa - Master's Thesis from Technical University of Denmark.

Tümmler B, Wiehlmann L, Klockgether J & Cramer N (2014) Advances in understanding Pseudomonas. *F1000Prime Rep.* **6:** 9

Varma A & Palsson B (1994) Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild-Type Escherichia coli W3110. **60:** 3724–3731

Waters CK (2007) The nature and context of exploratory experimentation: an introduction to three case studies of exploratory research. *Hist. Philos. Life Sci.* **29:** 275–284

Weigand MR & Sundin GW (2012) General and inducible hypermutation facilitate parallel adaptation in Pseudomonas aeruginosa despite divergent mutation spectra. *Proc. Natl. Acad. Sci. U. S. A.* **109:** 13680–5

Wiechert W (2001) 13C metabolic flux analysis. *Metab. Eng.* **3:** 195–206

Wodke JAH, Puchałka J, Lluch-Senar M, Marcos J, Yus E, Godinho M, Gutiérrez-Gallego R, dos Santos VAPM, Serrano L, Klipp E & Maier T (2013) Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.* **9:** 653

Worlitzsch D, Tarran R, Ulrich M, Schwab U, Cekici A, Meyer KC, Birrer P, Bellon G, Berger J, Weiss T, Botzenhart K, Yankaskas JR, Randell S & Boucher RC (2002) Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients. *J. Clin. Invest.* **109:** 317–325

Yang L, Jelsbak L, Marvig RL, Damkiær S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK, Ciofu O, Høiby N, Sommer MOA & Molin S (2011) Evolutionary dynamics of bacteria in a human host environment. *Proc. Natl. Acad. Sci. U. S. A.* **108:** 7481–6

Zhang WCCC, Shyh-Chang N, Yang H, Rai A, Umashankar S, Ma S, Soh BSSS, Sun LLLL, Tai BCCC, Nga MEEE, Bhakoo KKKK, Jayapal SRRR, Nichane M, Yu Q, Ahmed DAAA, Tan C, Sing WPPP, Tam J, Thirugananam A, Noghabi MSSS, et al (2012) Glycine Decarboxylase Activity Drives Non-Small Cell Lung Cancer Tumor-Initiating Cells and Tumorigenesis. *Cell* **148:** 259–272

Zur H, Ruppin E & Shlomi T (2010) iMAT: an integrative metabolic analysis tool. *Bioinformatics* **26:** 3140–2

Appendix A

The analysis of genome-scale metabolism in *P. aeruginosa* presented in **Paper 2** revealed a metabolic shift through the glycine cleavage system and surrounding reactions (Figure 5). We wanted to test if the metabolic subsystem depicted in Figure 5 was also subject to changes in other lineages of *Pseudomonas aeruginosa*.

We decided to look for mutations (missense, nonsense and indels) within the listed genes from Figure 5 among the recently published genome-sequences of 474 clinical *P. aeruginosa* isolates representing 36 different lineages (Marvig *et al*, 2015). A mutation was counted for each unique mutation that was found for at least one isolate within each lineage. The total number of unique mutations within the dataset was 6738 distributed on 5677 genes (Marvig *et al*, 2015). From the same dataset we found 32 mutations among the 19 genes from Figure 5. To test if 32 mutations among 19 genes were statistically different from what we would expect based on the frequency of 6738 mutations in 5677, we used a Poisson distribution test, where *x* is the number of counts and the mean number of counts is *λ* (Johnson, 2005):

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0,1,2,\ldots \ \lambda > 0$$

From the total data set we can calculate the frequency of mutations per gene, *α*:

$$\text{Mutations per gene: } \alpha = \frac{\text{total number of mutations}}{\text{total number of genes}} = \frac{6738}{5677} = 1.19$$

We can estimate *λ* for *T* genes (in this case 19 genes):

$$\text{Mean mutation frequency in 19 genes: } \lambda = T \cdot \alpha = 19 \cdot 1.19 = \ 22.6$$

Now we can use the Poisson distribution for calculating the probability of 32 mutations in 19 genes:

$$P(X \geq 32) = 1 - F(31,22.6) = \ 0.035$$

The probability of finding at least 32 mutations in 19 genes is therefore below 5% (p<0.05).

References:

Johnson RA (2005). Miller and Freund's probability and Statistics for Engineers 7th ed. Upper Saddle River, NJ: Pearson Prentice Hall

Marvig RL, Sommer LM, Molin S & Johansen HK (2015). Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nat. Genet*. **47**: 57–65